



Maximum-Likelihood Estimation and Scoring Under Parametric Constraints

by Terrence Moore and Brian Sadler

ARL-TR-3805

May 2006

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory

Adelphi, MD 20783-1197

ARL-TR-3805

May 2006

Maximum-Likelihood Estimation and Scoring Under Parametric Constraints

Terrence Moore and Brian Sadler
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
<p>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) May 2006		2. REPORT TYPE Final		3. DATES COVERED (From - To) FY05 to FY06	
4. TITLE AND SUBTITLE Maximum-Likelihood Estimation and Scoring Under Parametric Constraints				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Terrence Moore and Brian Sadler				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: AMSRD-ARL-CI-CN 2800 Powder Mill Road Adelphi, MD 20783-1197				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-TR-3805	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory 200 Powder Mill Road Adelphi, MD 20783-1197				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Maximum likelihood (ML) estimation is a popular approach in solving many signal processing problems. Many of these problems cannot be solved analytically and so numerical techniques such as the method of scoring are applied. However, in many scenarios, it is desirable to modify the ML problem with the inclusion of additional side information. Often this side information is in the form of parametric constraints which the ML estimate (MLE) must now satisfy. We examine the asymptotic normality of the constrained ML (CML) problem and show that it is still consistent as well as asymptotically efficient (with respect to the constrained Cramér-Rao bound). We also generalize the method of scoring to include the constraints, and satisfy the constraints after each iterate. Convergence properties and examples verify the usefulness of the constrained scoring approach. As a particular example, an alternative and more general CML estimator is developed for the linear model with linear constraints.					
15. SUBJECT TERMS Constraint estimation, maximum likelihood, scoring					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UL	18. NUMBER OF PAGES 51	19a. NAME OF RESPONSIBLE PERSON Terrence Moore
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER (Include area code) (301) 394-1236

Contents

1. Introduction	1
2. Preliminaries	2
2.1 Problem Statement and Definitions	3
2.2 The Constrained Cramer-Rao Lower Bound	4
3. Asymptotic Normality of the CMLE	6
4. Scoring with Constraints	11
4.1 The Projection Step	11
4.2 The Restoration Step	16
4.3 Implementation of the Constrained Scoring Algorithm	17
5. Convergence Properties	20
6. The Linear Model with Constraints	23
6.1 Linear Constraints	24
6.2 Nonlinear Constraints	27
7. A Nonlinear Model Example	31
7.1 MIMO Instantaneous Mixing Model	31
8. Conclusions	35
References	36
Appendices	39

A. Equivalence of Optimality Conditions	39
B. Taylor Expansion Derivation	41
C. Constrained Least Squares Estimator	43
Distribution	45

List of Figures

1	The constrained scoring algorithm.	18
2	The average norm of $\mathbf{x} - \mathcal{H}\boldsymbol{\vartheta}_k$ at iteration k . There is significant gain in the first iterate compared with later iterates, as expected with a quadratically convergent sequence.	29
3	Average mean-square error of the CSA compared with the CCRB.	30
4	Average bias error of the CSA.	30
5	Average MSE of the elements of (a) the two sources and (b) the two channels compared with the mean CCRB. Note the CCRB for the channels overlap.	34
6	The average decrement at each iteration.	34

List of Tables

1	Complexity of the CSA per iteration.	18
---	----------------------------------------------	----

1. Introduction

Maximum likelihood (ML) estimation is a popular approach in solving signal processing problems, especially in scenarios with a large data set, where the maximum likelihood estimator (MLE) is in many senses optimal due to its asymptotic characteristics. The procedure simply relies on maximizing the likelihood equation, and, in analytically intractable cases, the MLE can still be obtained iteratively through methods of optimization, e.g., via the method of scoring. However, in many signal processing problems, it is desirable or necessary to perform ML estimation when side information is available. Often this additional information is in the form of parametric equality or inequality constraints on a subset of the parameters. Examples of side information include the constant modulus property, some known signal values (semi-blind problems), restricted power levels (e.g., in networks), known angles of arrival, array calibration, some forms of precoding, and so on. With the addition of these parametric constraints, this procedure is now called constrained maximum likelihood (CML) estimation, with the solution being the constrained maximum likelihood estimator (CMLE).

As a measure of performance for the MLE, the Cramér-Rao bound (CRB), obtained via the inverse of the Fisher information matrix (FIM), is the lower bound of the error covariance of any unbiased estimator. However, it is desirable to measure performance of estimators that satisfy the side information constraints. Gorman and Hero used the Chapman-Robbins bound to develop a constrained version of the CRB which lower bounds the error covariance for constrained, unbiased estimators for the case when the unconstrained model has a nonsingular FIM (1). Marzetta simplified their derivation and formulation for the same nonsingular FIM case (2). Then, Stoica and Ng constructed a more general formulation of the CRB that incorporates the constraint information without the assumption of a full-rank FIM (3). Their constrained CRB (CCRB) was also shown to subsume the previous cases which require a nonsingular FIM.

The MLE is an optimal choice for an estimator in the sense that asymptotically the MLE is both consistent and efficient (4). For the case of linear equality constraints, Osborne showed that this result is preserved with the error covariance of the CMLE approaching the CCRB (5). Osborne's result was obtained independently and is further confirmation of the CCRB result in (3), although the CCRB is not discussed in (5). We extend this asymptotic normality result for the CMLE to the more general nonlinear constraint case. Specifically, we show that the CMLE is also consistent and asymptotically efficient with respect to the CCRB.

Although the ML problem is easy to express, obtaining the MLE is often a difficult task. Fortunately, iterative techniques, such as the method of scoring (4,6), are available which

reach the MLE under certain conditions. Typically, scoring is applied on top of an existing method which provides the initialization (7,8). However, those schemes must be adjusted when constraints have been added to the model. Prior research has focused on developing iterative techniques based on Lagrangian methods to obtain the CMLE (5,9,10). However, convergence properties and criteria are overlooked.

The constrained scoring algorithm (CSA) developed here is a generalization of the method of scoring to obtain the CMLE under certain conditions. The scheme relies on alternating between a projection step, similar to gradient based descent iterations, and a restorative step which ensures that the solution satisfies the parametric constraints. We, furthermore, detail several convergence properties associated with this CSA. Convergence of iterative techniques are always dependent on the initialization, but the results obtained here show that this method obtains at least a local MLE. Thus, with a sufficiently accurate initialization, the CSA will in fact obtain the CMLE.

We provide several examples to demonstrate the effectiveness of the CSA. First, we examine the classical CMLE problem when imposing linear constraints on a linear model. The CSA analytically solves this problem in a single step and provides an equivalent alternative to the traditional answer (e.g., see (4, p. 252) and (11, p. 299)), where the new solution is applicable under weaker conditions. We also demonstrate that our CMLE and the traditional solution are both unbiased and efficient. Second, we find the CMLE after imposing nonlinear, nonconvex constraints on the signal modulus in a complex-valued linear model. Provided the initialization is sufficiently close, our simulations show evidence of unbiasedness and efficiency for this particular choice of constraint as well. Third, we consider a nonlinear model parameterization from (12). In this case, constant modulus and semiblind constraints are applied on the signal that passes through an instantaneous mixing channel.

This paper is developed as follows: In the following section, we formally state the CML problem and give definitions for necessary terms used throughout. In section 3, we determine the asymptotic normality properties of the CMLE, showing both consistency and asymptotic efficiency. Next, in section 4, we develop the CSA via the method of Lagrange multipliers and natural projections. In section 5, we discuss the convergence properties of the given CSA. In sections 6 and 7, we provide some examples that illustrate the effectiveness of the CSA.

2. Preliminaries

In this and subsequent sections, we use the following notation: Scalars will be in lowercase, vectors in bold font, and matrices in uppercase bold font (e.g., a is a scalar, \mathbf{a} a vector, and

\mathbf{A} a matrix), where it is understood that $(\cdot)_i$ will denote the i th element of a vector or row/column of a matrix and $(\cdot)_{ij}$ will denote the i th row, j th column element of a matrix. We will denote $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H$ as the transpose, the conjugate and the conjugate transpose, respectively, of either a vector or matrix. For matrices, $(\cdot)^{-1}$ is the matrix inverse and $(\cdot)^\dagger$ the pseudoinverse. All vectors will be column vectors. Sets will be denoted in capital Greek letters, and all sets will be a subset of a Euclidean metric space, i.e., $(\mathbb{R}^n, \|\cdot\|)$ for some positive integer n and where $\|\cdot\|$ is the L^2 -norm. When applied to matrices, $\|\cdot\|$ will be the Frobenius norm.

2.1 Problem Statement and Definitions

We have a vector of observations \mathbf{x} in a sample space $\Omega \subset \mathbb{R}^M$ satisfying the likelihood function of a known form $p(\mathbf{x}; \boldsymbol{\theta})$. We want to estimate the unknown $\boldsymbol{\theta}$ parameter vector under the assumption that $\boldsymbol{\theta}$ is restricted to a closed, convex set $\Theta \subset \mathbb{R}^N$, which we will assume can be defined parametrically.¹ Let $\boldsymbol{\theta}_o \in \Theta$ be the true vector of parameters. Then the CMLE $\hat{\boldsymbol{\theta}}(x)$ is given by

$$\hat{\boldsymbol{\theta}}(x) = \arg \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}; \boldsymbol{\theta}). \quad (1)$$

Since $-\log(\cdot)$ is strictly monotone decreasing, this CMLE can alternately be viewed as the solution to the following constrained optimization problem

$$\min_{\boldsymbol{\theta}} \quad -\log p(\mathbf{x}; \boldsymbol{\theta}) \quad (2)$$

$$\text{s.t.} \quad \mathbf{f}(\boldsymbol{\theta}) = \mathbf{0} \quad (3)$$

$$\mathbf{g}(\boldsymbol{\theta}) \leq \mathbf{0} \quad (4)$$

where the the negative log-likelihood function is the objective function, and $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^K$ and $\mathbf{g} : \mathbb{R}^N \rightarrow \mathbb{R}^L$ are the functional constraints which define the constraint set, i.e., $\Theta = \{\boldsymbol{\theta} : \mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}, \mathbf{g}(\boldsymbol{\theta}) \leq \mathbf{0}\}$. We make the assumption that $p(\mathbf{x}; \boldsymbol{\theta})$, $\mathbf{f}(\boldsymbol{\theta})$, and $\mathbf{g}(\boldsymbol{\theta})$ all have continuous second derivatives with respect to $\boldsymbol{\theta}$. Derivatives will be expressed either as $\nabla_{\boldsymbol{\theta}}(\cdot)$ or as $\frac{\partial}{\partial \boldsymbol{\theta}}(\cdot)$. We also require that the functional constraints be consistent, i.e., $\Theta \neq \emptyset$. All expectations will be with respect to the appropriate distribution of the likelihood, i.e. $E_{\boldsymbol{\theta}}(\cdot) = \int_{\mathbf{y} \in \Omega} (\cdot) p(\mathbf{y}; \boldsymbol{\theta}) d\mathbf{y}$.

A point $\boldsymbol{\theta}$ which satisfies the functional constraints is said to be *feasible*, and Θ is referred to as the *feasible region*. The i th inequality constraint $g_i(\boldsymbol{\theta}) \leq 0$ is said to be *active* at a feasible point $\boldsymbol{\theta}$ if $g_i(\boldsymbol{\theta}) = 0$, otherwise it is *inactive*. Thus, the equality constraints $\mathbf{f}(\boldsymbol{\theta})$ are always considered active. A feasible point $\boldsymbol{\theta}$ is a *regular* point of the constraints in \mathbf{f} and the active constraints in \mathbf{g} if the vectors $\nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{\theta})$, $\nabla_{\boldsymbol{\theta}} g_j(\boldsymbol{\theta})$, $1 \leq i \leq K$, $1 \leq j \leq L'$, are

¹The condition that Θ be convex is not so much a strict requirement as it is a convenient one. More discussion on this issue is found in section 4.3.

linearly independent, where we assume only the first L' constraints of \mathbf{g} are active. Thus, a regular point requires no redundancy in the active constraints. Properties for these terms can be found in (13,14).

Define the gradient matrices of \mathbf{f} and \mathbf{g} by the continuous functions

$$\mathbf{F}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^T \mathbf{f}(\boldsymbol{\theta}) \text{ and } \mathbf{G}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^T \mathbf{g}(\boldsymbol{\theta}).$$

Note that, assuming $\boldsymbol{\theta}$ is regular in the active constraint set, $\mathbf{F}(\boldsymbol{\theta})$ has full row rank K , whereas $\mathbf{G}(\boldsymbol{\theta})$ is not necessarily so. Define $\mathbf{U} : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times J}$ to be a continuous function such that for each $\boldsymbol{\theta}$, $\mathbf{U}(\boldsymbol{\theta})$ is a matrix whose columns form an orthonormal null space of the range space of the row vectors in $\mathbf{F}(\boldsymbol{\theta})$, i.e., such that

$$\mathbf{F}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0} \text{ and } \mathbf{U}^T(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}) = \mathbf{I}_{J \times J} \quad (5)$$

for every $\boldsymbol{\theta} \in \mathbb{R}^N$. If $\boldsymbol{\theta}$ is regular then $\mathbf{U} : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times N-K}$, i.e., then $J = N - K$ since $\mathbf{F}(\boldsymbol{\theta})$ is full row rank. Also, note that \mathbf{U} is independent of $\boldsymbol{\theta}$ whenever \mathbf{F} is. This occurs in the linearly constrained case, when $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{F}\boldsymbol{\theta} + \mathbf{v}$ for some matrix $\mathbf{F} \in \mathbb{R}^{N \times N}$ and vector $\mathbf{v} \in \mathbb{R}^N$. So, the gradient $\mathbf{F}(\boldsymbol{\theta})$ is a constant \mathbf{F} and, thus, $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{U}$.

2.2 The Constrained Cramer-Rao Lower Bound

The error covariance of any unbiased estimator $\bar{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$ is bounded by the Cramer-Rao Lower Bound (CRB). The classical development of the Cramer-Rao Lower Bound (CRB) is well-known (4,15). The bound is expressed as

$$E_{\boldsymbol{\theta}}((\bar{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta})^T) \geq \mathbf{I}^{-1}(\boldsymbol{\theta}) \quad (6)$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the FIM given by

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}. \quad (7)$$

The CRB and FIM exist provided the following *regularity* conditions² hold:

$$E_{\boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \text{ and } \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} \bar{\boldsymbol{\theta}}(\mathbf{x}) = E_{\boldsymbol{\theta}} \bar{\boldsymbol{\theta}}(\mathbf{x}) \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \quad (8)$$

²In the statistical inference literature a point $\boldsymbol{\theta}$ that satisfies these regularity conditions is sometimes said to be *regular*, or *information-regular* (16). To avoid the confusion between that definition and the optimization literature's definition in the prior subsection, all instances of *regular* in this paper will be in the optimization context and all instances of *regularity* will be in the statistical inference context.

Note these must hold for all $\boldsymbol{\theta}$. Additionally, if we also have that

$$E_{\boldsymbol{\theta}} \frac{\partial^2 \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}, \quad (9)$$

then we can use an alternate expression for the FIM:

$$\mathbf{I}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}} \frac{\partial^2 \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \quad (10)$$

From equation 6, the existence of the CRB requires a nonsingular FIM as well. When the model is not locally identifiable, however, then the FIM is singular and the CRB is not always defined.³ To obtain a CRB, the model must include sufficient constraints on the parameters to achieve identifiability and, hence, a nonsingular FIM. The choice of constraints are completely dictated by the model, and, thus, applying such constraints previously required a reevaluation of the FIM on some appropriate dimension-reducing reparameterization of $\boldsymbol{\theta}$ that includes this constraint information. A bound is then obtained by a transformation from the reparameterization back to $\boldsymbol{\theta}$. Since the reevaluated FIM depends on the reparameterization, i.e., it needs to be evaluated for every reparameterization or unique set of constraints, the classical Fisher Information theory does not incorporate the constraint information in any convenient closed form. Even when the original model is identifiable, the classical theory ignores the contribution of the side information in the form of a specified constraint set.

To overcome this deficiency, Gorman and Hero (1) and then Marzetta (2) developed formulations of the CRB which do include constraint information for the case where the FIM is full-rank. Improving on their work, Stoica and Ng (3) formulated a CCRB that explicitly incorporates the active constraint information with the original FIM, singular or nonsingular.

Theorem 1 (Stoica & Ng (3)). *Assume we know the active constraints and that these are all incorporated into \mathbf{f} . Let $\bar{\boldsymbol{\theta}}(\mathbf{x})$ be an unbiased estimate of $\boldsymbol{\theta}$ satisfying the active functional constraints in equation 3. Then, under certain regularity conditions, if $\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})$ is nonsingular,*

$$E_{\boldsymbol{\theta}}((\bar{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta})(\bar{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta})^T) \geq \mathbf{U}(\boldsymbol{\theta})(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}\mathbf{U}^T(\boldsymbol{\theta}) \triangleq \mathbf{B}(\boldsymbol{\theta}) \quad (11)$$

where equality is achieved if and only if (in the mean square sense)

$$\bar{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta} = \mathbf{U}(\boldsymbol{\theta})(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}\mathbf{U}^T(\boldsymbol{\theta})\nabla \log p(\mathbf{x}; \boldsymbol{\theta}).$$

In the evaluation of this CCRB, the inactive inequality constraints do not contribute any side information. This is so because we made the assumption that the inequality

³Often, the pseudoinverse $\mathbf{I}^\dagger(\boldsymbol{\theta})$ is used, but this bound is not always valid except under certain conditions (17).

constraints are inactive (1), i.e., only active constraints affect the outcome of the estimator. Also note that rather than requiring a non-singular FIM $\mathbf{I}(\boldsymbol{\theta})$, the alternate condition is that $\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})$ be non-singular. Thus, the unconstrained FIM may still be singular, or, equivalently, the unconstrained model unidentifiable, but the constrained model must be identifiable, at least locally.

The natural extensions to the classical model, e.g., the CRB on differentiable transformations (4, p. 45) and the CRB for biased bounds, also can be applied to this constrained formulation (12).

3. Asymptotic Normality of the CMLE

Asymptotic properties of the MLE can be found in (4). This motivates the desire to obtain corresponding results for the CMLE. In particular, we wish to show results on asymptotic consistency and efficiency for the constrained case. For asymptotic consistency, we will rely on the Kullback-Leibler information (4). For asymptotic efficiency, since the constrained maximum likelihood problem equation 1 is equivalent to the constrained optimization problem equations 2 through 4, we will use the tools of optimization theory. The main results of this section, equations 18 and 32, can be summarized as follows.

Theorem 2. *Assuming the pdf $p(\mathbf{x}; \boldsymbol{\theta})$ satisfies certain regularity conditions, e.g., as in Theorem 1, then the CMLE $\hat{\boldsymbol{\theta}}_n$ is asymptotically distributed according to*

$$\hat{\boldsymbol{\theta}}_n \sim \mathcal{N}(\boldsymbol{\theta}_o, \mathbf{B}(\boldsymbol{\theta}_o)) \quad (12)$$

where $\boldsymbol{\theta}_o$ is the true parameter vector.

Proof. Suppose we observe the data set $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ where each \mathbf{x}_i is independently distributed as \mathbf{x} , i.e., we observe n iid samples from a distribution of the known form $p(\mathbf{x}, \boldsymbol{\theta})$, in order to estimate $\boldsymbol{\theta}$. And, again, $\boldsymbol{\theta}_o$ is the true parameter vector. For this section, denote $\hat{\boldsymbol{\theta}}_n(\mathbf{x}_1, \dots, \mathbf{x}_n)$ as the CMLE based on n samples. For convenience, in this section, the CMLE will also often be denoted $\hat{\boldsymbol{\theta}}_n$. Then, the CMLE on n samples maximizes the joint density of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$, i.e.,

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}_1, \dots, \mathbf{x}_n) = \arg \max_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^n p(\mathbf{x}_i, \boldsymbol{\theta}). \quad (13)$$

Or, equivalently, since log is monotone,

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}_1, \dots, \mathbf{x}_n) = \arg \max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \log \prod_{i=1}^n p(\mathbf{x}_i, \boldsymbol{\theta}) \quad (14)$$

$$= \arg \max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \log p(\mathbf{x}_i, \boldsymbol{\theta}). \quad (15)$$

As $n \rightarrow \infty$, by the law of large numbers we have that

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}_1, \dots, \mathbf{x}_n) \rightarrow \arg \max_{\boldsymbol{\theta} \in \Theta} E_{\boldsymbol{\theta}_o} \log p(\mathbf{x}; \boldsymbol{\theta}) \quad (16)$$

The Kullback-Liebler information satisfies

$$E_{\boldsymbol{\theta}} \log \frac{p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\phi})} = \int_{\Omega} \log \frac{p(\mathbf{x}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\phi})} p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \geq 0 \quad (17)$$

with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\phi}$. Thus, we have that $E_{\boldsymbol{\theta}_o} \log p(\mathbf{x}; \boldsymbol{\theta}) \leq E_{\boldsymbol{\theta}_o} \log p(\mathbf{x}; \boldsymbol{\theta}_o)$ with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_o$. Thus, the solution of the maximization in equation 16 is $\boldsymbol{\theta}_o$. That is, as $n \rightarrow \infty$ then

$$\hat{\boldsymbol{\theta}}_n(\mathbf{x}_1, \dots, \mathbf{x}_n) \rightarrow \boldsymbol{\theta}_o, \quad (18)$$

and the CMLE is consistent.

Next we determine the asymptotic covariance characteristics of $\hat{\boldsymbol{\theta}}_n$ employing tools from optimization theory. One such useful tool in converting the constrained optimization problem equation 2 into an unconstrained problem is the method of Lagrange multipliers. The method develops a *Lagrangian* function which incorporates the objective and constraints so that solutions of the optimization problem must be stationary points of this function.⁴ The Lagrangian of equation 2 is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = -\log p(\mathbf{x}; \boldsymbol{\theta}) + \boldsymbol{\mu}^T \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\nu}^T \mathbf{g}(\boldsymbol{\theta}). \quad (19)$$

The vectors $\boldsymbol{\mu} \in \mathbb{R}^K$ and $\boldsymbol{\nu} \in \mathbb{R}^L$ are the *Lagrange multipliers* of the function. Any potential solution of equation 2 must be a *stationary point* of equation 19, i.e., it must be a point $\boldsymbol{\theta}^*$ satisfying the following Karush-Kuhn-Tucker (KKT) necessary conditions (18, p.243):

$$\nabla_{\boldsymbol{\mu}}^T \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = \mathbf{f}(\boldsymbol{\theta}^*) = \mathbf{0} \quad (20)$$

$$\mathbf{g}(\boldsymbol{\theta}^*) \leq \mathbf{0} \quad (21)$$

$$\boldsymbol{\nu}^* \geq \mathbf{0} \quad (22)$$

$$\boldsymbol{\nu}^{*T} \mathbf{g}(\boldsymbol{\theta}^*) = \mathbf{0} \quad (23)$$

$$\nabla_{\boldsymbol{\theta}}^T \mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = -\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*) + \boldsymbol{\mu}^{*T} \mathbf{F}(\boldsymbol{\theta}^*) + \boldsymbol{\nu}^{*T} \mathbf{G}(\boldsymbol{\theta}^*) = \mathbf{0}. \quad (24)$$

Note the first two conditions, equations 20 and 21, are simply the constraints equations 3 and 4 from the constrained optimization problem. Also, equation 23 implies that ν_i^* is nonzero if and only if the constraint g_i is active. Either ν_i^* or $g_i(\boldsymbol{\theta}^*)$ is zero, exclusively, so

⁴A commonly used alternative optimality condition for convex sets is discussed in appendix A.

equations 21 through 23 actually define L explicit equations. Since equation 20 defines K equations and equation 24 defines N equations, this system above contains exactly $K + L + N$ equations with $K + L + N$ unknowns $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$. Where these equations are not redundant, i.e., at a regular point, the solution $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$ is locally unique. This solution $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$, however, is typically not analytically tractable and requires a numerical optimization approach.

The last equation can be conveniently simplified by considering only the active constraints at a stationary point $\boldsymbol{\theta}^*$, for similar reasons given in the CCRB development. So for any stationary point, we will assume that all the active constraints are already incorporated into \mathbf{f} and modify \mathbf{F} and \mathbf{U} accordingly. Hence, we can rewrite equation 24 as

$$-\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*) + \boldsymbol{\mu}^{*T} \mathbf{F}(\boldsymbol{\theta}^*) = \mathbf{0}. \quad (25)$$

This implies that the gradient of the log-likelihood is in the range space of the gradient of the active equality constraints at the stationary point. Hence, geometrically, the direction of steepest descent of the objective function must be orthogonal to the tangent plane of \mathbf{f} at stationary points of the Lagrangian. And since $\hat{\boldsymbol{\theta}}_n$ is a stationary point, equations 20 through 25 hold at $\hat{\boldsymbol{\theta}}_n$. From equation 5, $\mathbf{F}(\hat{\boldsymbol{\theta}}_n)\mathbf{U}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$, so equation 25 implies that a necessary condition for $\hat{\boldsymbol{\theta}}_n$ to be the CMLE is for

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \hat{\boldsymbol{\theta}}_n) \mathbf{U}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}. \quad (26)$$

For a moment, it will be convenient to consider the the equivalent condition of equation 26 in vector notation, i.e., for $1 \leq j \leq J$,

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \hat{\boldsymbol{\theta}}_n) \mathbf{u}_j(\hat{\boldsymbol{\theta}}_n) = 0 \quad (27)$$

where $\mathbf{u}_j(\boldsymbol{\theta})$ is the j th column of $\mathbf{U}(\boldsymbol{\theta})$. Now let $\boldsymbol{\theta}$ be a point near $\hat{\boldsymbol{\theta}}_n$. The Taylor expansion (19) of equation 27 about $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}_n$ gives us

$$\begin{aligned} 0 &= \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}) \mathbf{u}_j(\boldsymbol{\theta}) + \mathbf{u}_j^T(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ &\quad + \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \mathbf{u}_j(\boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o(1) \\ &= \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}) [\mathbf{u}_j(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}^T \mathbf{u}_j(\boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o(1)] \\ &\quad + \mathbf{u}_j^T(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o(1) \\ &= \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}) \mathbf{u}_j(\hat{\boldsymbol{\theta}}_n) + \mathbf{u}_j^T(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}) (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o(1) \end{aligned} \quad (28)$$

where the $o(1)$ term is the sum of the higher order terms of $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$, which vanishes as $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \rightarrow 0$. Collecting the vectors in equation 28 in matrix notation, we have

$$\mathbf{0} = \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) + \mathbf{U}^T(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o(1).^5 \quad (29)$$

Since the CMLE is consistent from equation 18, then for sufficiently large n , the CMLE is close to the true parameter vector $\boldsymbol{\theta}_o$, and so the equation is satisfied for the true parameter vector $\boldsymbol{\theta}_o$, i.e., when $\boldsymbol{\theta} = \boldsymbol{\theta}_o$. So,

$$\mathbf{0} = \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) + \mathbf{U}^T(\boldsymbol{\theta}_o) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) + o(1). \quad (30)$$

Recall that $\boldsymbol{\theta}_o$ is also subject to the active (equality) constraints, i.e., it is a feasible point just as $\hat{\boldsymbol{\theta}}_n$ is. Since $\boldsymbol{\Theta}$ is connected, there exists a path-connected curve on the surface of $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}$ including the points $\{\hat{\boldsymbol{\theta}}_n : n = 1, 2, \dots\}$ and $\boldsymbol{\theta}_o$. Such a curve, in the optimization context, is called a *feasible arc* since every point on the curve satisfies the equality constraints. Let the continuously differentiable map $\tilde{\boldsymbol{\theta}} : \mathbb{R} \rightarrow \boldsymbol{\Theta}$ denote this feasible arc such that $\tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}_o$ and $\tilde{\boldsymbol{\theta}}(\frac{1}{n}) = \hat{\boldsymbol{\theta}}_n$ for each n . This arc reflects the consistency result since $\hat{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}(\frac{1}{n}) \rightarrow \tilde{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}_o$ as $n \rightarrow \infty$. Since $\mathbf{f}(\tilde{\boldsymbol{\theta}}(t)) = \mathbf{0}$ for all t , then

$$\mathbf{0} = \left. \frac{d}{dt} \mathbf{f}(\tilde{\boldsymbol{\theta}}(t)) \right|_{t=0} = \nabla_{\boldsymbol{\theta}}^T \mathbf{f}(\tilde{\boldsymbol{\theta}}(t)) \left. \frac{d}{dt} \tilde{\boldsymbol{\theta}}(t) \right|_{t=0} = \mathbf{F}(\boldsymbol{\theta}_o) \left. \frac{d}{dt} \tilde{\boldsymbol{\theta}}(t) \right|_{t=0} = \mathbf{F}(\boldsymbol{\theta}_o) \left. \frac{d}{dt} \tilde{\boldsymbol{\theta}}(t) \right|_{t=0}.$$

Thus, from equation 5, $\left. \frac{d}{dt} \tilde{\boldsymbol{\theta}}(t) \right|_{t=0} \in \text{span } \mathbf{U}(\boldsymbol{\theta}_o)$. Hence, using the Lagrange remainder form for the Taylor series (19), we have that

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o = \tilde{\boldsymbol{\theta}}(\frac{1}{n}) - \tilde{\boldsymbol{\theta}}(0) = \frac{1}{n} \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \mathbf{q}_n \quad (31)$$

for some $0 < s(n) < \frac{1}{n}$ and some $\mathbf{q}_n \in \mathbb{R}^J$. Substituting equation 31 into 30, we have

$$\mathbf{0} = \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) - \mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \mathbf{q}_n + o(1).$$

Given the regularity condition of Theorem 1 at $\boldsymbol{\theta}_o$, then for sufficiently large n we have that the matrix $\mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n)))$ is invertible for $\tilde{\boldsymbol{\theta}}$ in a neighborhood of $\boldsymbol{\theta}_o$. Therefore, solving for \mathbf{q}_n we find that

$$\mathbf{q}_n = \left(\mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \right)^{-1} \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) + o(1).$$

⁵By $o(1)$ we mean a vector which has all its elements $o(1)$.

Substituting \mathbf{q}_n back into equation 31 yields

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o = \frac{1}{n} \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \left(\mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \right)^{-1} \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) + \mathbf{o}(1),$$

or

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o) = \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \left(\mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \right)^{-1} \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) + \mathbf{o}(1).$$

Now, let $n \rightarrow \infty$. Then $\hat{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}_o$ by asymptotic consistency from equation 18, $\mathbf{o}(1) \rightarrow \mathbf{0}$, and $-\frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \rightarrow \mathbf{I}(\boldsymbol{\theta}_o)$ by the law of large numbers, i.e., since $\frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_o))$ (4). Also $\mathbf{U}(\hat{\boldsymbol{\theta}}_n) \rightarrow \mathbf{U}(\boldsymbol{\theta}_o)$ by continuity, and $\mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \rightarrow \mathbf{U}(\boldsymbol{\theta}_o)$ by the pinching theorem and continuity. Thus, we have that as $n \rightarrow \infty$,

$$\begin{aligned} \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \left(\mathbf{U}^T(\boldsymbol{\theta}_o) \frac{1}{n} \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_o) \mathbf{U}(\tilde{\boldsymbol{\theta}}(s(n))) \right)^{-1} \mathbf{U}^T(\hat{\boldsymbol{\theta}}_n) \\ \rightarrow \mathbf{U}(\boldsymbol{\theta}_o) (\mathbf{U}^T(\boldsymbol{\theta}_o) \mathbf{I}(\boldsymbol{\theta}_o) \mathbf{U}(\boldsymbol{\theta}_o))^{-1} \mathbf{U}^T(\boldsymbol{\theta}_o) \text{ a.s.} \end{aligned}$$

From the asymptotic normality result on the MLE (4, 15), $\text{Cov}(\frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_o)) \rightarrow \mathbf{I}(\boldsymbol{\theta}_o)$ as $n \rightarrow \infty$ where the convergence is convergence in distribution. And thus, by Slutsky's theorem (15), we have that

$$\text{Cov}_{\boldsymbol{\theta}_o}(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_o)) \rightarrow \mathbf{U}(\boldsymbol{\theta}_o) (\mathbf{U}^T(\boldsymbol{\theta}_o) \mathbf{I}(\boldsymbol{\theta}_o) \mathbf{U}(\boldsymbol{\theta}_o))^{-1} \mathbf{U}^T(\boldsymbol{\theta}_o) = \mathbf{B}(\boldsymbol{\theta}_o) \quad (32)$$

where the convergence is in distribution. Since this is the CCRB in equation 11, this shows that the CMLE is asymptotically efficient. The result in equation 18, combined with equation 32, prove the theorem. \square

Theorem 2 is parallel to the classical asymptotic normality property for the classical (unconstrained) maximum likelihood estimate (4). And, this serves as another verification of the CCRB result in equation 11. In fact, in the absence of constraints, then $\mathbf{U}(\boldsymbol{\theta}_o) = \mathbf{I}_{N \times N}$ since \mathbf{f} is null, and provided the FIM is nonsingular, we have the classical asymptotic MLE result as well. The result in equation 12 was earlier shown by Osborne (5), but only for linear constraints, and no link was made regarding the constrained CRB.

4. Scoring with Constraints

An analytic, closed-form solution of the MLE is sometimes found from the first order conditions on the log-likelihood (the KKT equations equations 20 through 24 with null constraint functions), i.e., by solving for $\hat{\boldsymbol{\theta}}$ in

$$\left. \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0.$$

Solutions are the stationary points, and if a unique solution exists, it is the MLE. Similarly, a solution for the CMLE can be found by solving for the zero point of the arc $\tilde{\boldsymbol{\theta}}(t)$ satisfying

$$\left. \frac{d}{dt} \log p(\mathbf{x}; \tilde{\boldsymbol{\theta}}(t)) \right|_{t=0} = 0$$

for any feasible arc $\tilde{\boldsymbol{\theta}} : \mathbb{R} \rightarrow \boldsymbol{\Theta}$ parameterized by t . Again, if a unique solution exists, $\tilde{\boldsymbol{\theta}}(0)$ is the CMLE.

In general, however, analytic solutions of the MLE and CMLE problem are unavailable. This motivates the use of iterative procedures to attain the CMLE. In this section, we derive a generalization of the classical scoring algorithm by incorporating the side information contained in the constraints. This new method will include a *projection step* and a *restoration step* to ensure that each of the iterates remains both feasible and usable.

4.1 The Projection Step

The simplest iterative schemes are gradient methods of the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \cdot \mathbf{d}_k, \tag{33}$$

where $\alpha_k > 0$ and $\mathbf{d}_k \in \mathbb{R}^N$ are suitably chosen sequences of step sizes and step directions. Before continuing, it is appropriate to define a few terms relating to such iterative schemes (13,14,20,21,22). Given a feasible point $\boldsymbol{\theta}$, a *feasible step* consists of a step direction \mathbf{d} and step size α such that $\boldsymbol{\theta} + \alpha \mathbf{d}$ is also a feasible point. Thus, by suitably chosen, we mean in part that the sequences α_k and \mathbf{d}_k are chosen such that $\boldsymbol{\theta}_k$ is feasible for every positive integer k . However, not every sequence of feasible steps results in a sequence of feasible points which converges to a (local) minimum of the objective function. Thus, a *usable step* consists of a step direction \mathbf{d} and step size α such that the corresponding evaluation of the objective is less than that of the previous iterate, i.e.,

$$-\log p(\mathbf{x}; \boldsymbol{\theta}) > -\log p(\mathbf{x}; \boldsymbol{\theta} + \alpha \mathbf{d}).$$

So, a suitably chosen sequence of step sizes and step directions is a sequence of feasible and usable steps.

Of particular importance in the class of gradient methods is the subclass of Quasi-Newton methods of the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha_k \mathbf{D}_k \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k)$$

where \mathbf{D}_k is a sequence of positive definite symmetric matrices. In general, the gradient is of the objective function, which in our case is the negative log-likelihood function. Since the gradient of the objective function is the direction of steepest ascent, it's negative is the direction of steepest descent. The matrix \mathbf{D}_k then projects this direction vector to some purpose, dictated by the model. This class of methods includes

1. the method of steepest descent, where $\mathbf{D}_k = \mathbf{I}_{N \times N}$,
2. Newton's method, where $\mathbf{D}_k = (\nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_k))^{-1}$, as well as,
3. the method of scoring, where $\mathbf{D}_k = -\mathbf{I}(\boldsymbol{\theta}_k)$.

The method of steepest descent often leads to slow, linear convergence rates. Newton's method uses a quadratic approximation to converge quadratically, and thus faster, but requires a second-order differentiation. The method of scoring asymptotically stabilizes the iteration statistically by exchanging the Hessian of the objective with it's expected value, the negative FIM. As evident in the equalities in equations 17 and 10, this also removes the need for computing a second-order derivative, albeit in exchange with the required evaluation of an expectation (integral). None of these methods, however, consider the information and restrictions from the constraints.

To incorporate constraints⁶, we return to the necessary conditions for a stationary point of the Lagrangian, in particular equations 25 and 20,

$$-\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x})) + \hat{\boldsymbol{\mu}}^T(\mathbf{x}) \mathbf{F}(\hat{\boldsymbol{\theta}}(\mathbf{x})) = \mathbf{0} \quad (34)$$

$$\mathbf{f}(\hat{\boldsymbol{\theta}}(\mathbf{x})) = \mathbf{0}. \quad (35)$$

Again, if $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is regular then the above equations completely determine the CMLE; however, the solution is difficult to obtain analytically. Let $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ be a given point near the CMLE $\hat{\boldsymbol{\theta}}(\mathbf{x})$; so $\boldsymbol{\theta}_1$ is a sufficiently close approximation of $\hat{\boldsymbol{\theta}}(\mathbf{x})$ which also satisfies the

⁶There are numerous ways to incorporate these constraints, e.g., using a directional Taylor approximation of the likelihood as discussed in appendix B, applying the method of Newton elimination, or using a general Taylor approximation of the Lagrangian optimality condition as discussed here.

constraints. Now, consider the Taylor approximations of equations 34 and 35 about θ_1 evaluated at a nearby point θ :

$$-\nabla_{\theta} \log p(\mathbf{x}; \theta_1) - \nabla_{\theta}^2 \log p(\mathbf{x}; \theta_1)(\theta - \theta_1) + \mathbf{F}^T(\theta)\boldsymbol{\mu} = \mathbf{0} \quad (36)$$

$$\mathbf{f}(\theta_1) + \mathbf{F}(\theta_1)(\theta - \theta_1) = \mathbf{0}. \quad (37)$$

Since $\hat{\theta}(\mathbf{x})$ is locally unique in equations 34 and 35, then by dropping the higher order terms and yet still forcing the equality with $\mathbf{0}$, θ would be a closer point to $\hat{\theta}(\mathbf{x})$ than θ_1 is. But we still need to solve for θ to obtain this better approximation. To add to this difficulty, $\boldsymbol{\mu}$ is also unknown since it is an approximation of $\hat{\boldsymbol{\mu}}(\mathbf{x})$ which is also unknown.

In matrix form, equations 36 and 37 are

$$\begin{bmatrix} -\nabla_{\theta}^2 \log p(\mathbf{x}; \theta_1) & \mathbf{F}^T(\theta_1) \\ \mathbf{F}(\theta_1) & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \theta - \theta_1 \\ \boldsymbol{\mu} - \boldsymbol{\mu}_1 \end{bmatrix} = \begin{bmatrix} \nabla_{\theta} \log p(\mathbf{x}; \theta_1) - \mathbf{F}^T(\theta_1)\boldsymbol{\mu}_1 \\ -\mathbf{f}(\theta_1) \end{bmatrix} \quad (38)$$

where we exchanged a term in equation 36 via the first order approximation $\mathbf{F}^T(\theta_1)\boldsymbol{\mu} \approx \mathbf{F}^T(\theta)\boldsymbol{\mu}$ for small $\boldsymbol{\mu}$ (9) and added $-\mathbf{F}^T(\theta_1)\boldsymbol{\mu}_1$ to both sides, where $\boldsymbol{\mu}_1$ is a chosen initialization. (The choice of $\boldsymbol{\mu}_1$ will be irrelevant to our end result.) The matrix is commonly referred to as the KKT matrix, and the system as a KKT system (18). By approximating the negative Hessian with the FIM, we have

$$\begin{bmatrix} \mathbf{I}(\theta_1) & \mathbf{F}^T(\theta_1) \\ \mathbf{F}(\theta_1) & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} \theta - \theta_1 \\ \boldsymbol{\mu} - \boldsymbol{\mu}_1 \end{bmatrix} = \begin{bmatrix} \nabla_{\theta} \log p(\mathbf{x}; \theta_1) - \mathbf{F}^T(\theta_1)\boldsymbol{\mu}_1 \\ -\mathbf{f}(\theta_1) \end{bmatrix}. \quad (39)$$

So, if either the KKT matrix is nonsingular, or θ_1 is regular (so that $\mathbf{F}(\theta_1)$ has full row rank) and the FIM is nonsingular, then one of the coefficient matrices on the LHS of equations 38 and 39 must be nonsingular. Premultiplying equation 38 or 39 by the inverse of their respective coefficient matrix, if it exists, and solving for θ and $\boldsymbol{\mu}$ leads to an iterative scheme where $(\theta, \boldsymbol{\mu})$ are the updates of the given $(\theta_1, \boldsymbol{\mu}_1)$. Such a scheme iteratively estimates the desired parameter estimate $\hat{\theta}(\mathbf{x})$ as well as the accompanying Lagrange multipliers $\hat{\boldsymbol{\mu}}(\mathbf{x})$ which satisfies equation 34. Such a method is presented in (9,10).⁷ We do not desire such a method, since, in addition to increasing the number of parameters needed to be estimated, the solution $(\hat{\theta}(\mathbf{x}), \hat{\boldsymbol{\mu}}(\mathbf{x}), \hat{\nu}(\mathbf{x}))$ may not be locally unique, hampering convergence stopping criteria, even when the solution of $\hat{\theta}(\mathbf{x})$ is unique.

Note that if the constraints \mathbf{f} are linear and the negative log-likelihood is quadratic, then there are no higher order terms and equations 40 and 41 (as well as equations 36 and 37)

⁷It is interesting to note that these authors use a formula (9, p.823) in their scheme that was later found by Marzetta to be a variation of the CCRB formula applicable when the FIM for the unconstrained model is invertible (2).

do not give approximations, but are exact, i.e., then $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{x})$. Thus, solving for $\boldsymbol{\theta}$ given $\boldsymbol{\theta}_1$ can be done in one step in this case (see section 6.1).

However, if $\boldsymbol{\theta}_1$ is not a regular point then the constraints include redundancy and the coefficient matrices are not invertible (as can be seen from their Schur complements). Likewise, if the FIM $\mathbf{I}(\boldsymbol{\theta}_1)$ is singular, then the statistical KKT matrix in equation 39 is not invertible. So it is desirable to find a scheme which does not rely on inverting these possibly singular matrices. We can rewrite equation 39 as

$$\mathbf{I}(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) + \mathbf{F}^T(\boldsymbol{\theta}_1)(\boldsymbol{\mu} - \boldsymbol{\mu}_1) = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1) - \mathbf{F}^T(\boldsymbol{\theta}_1)\boldsymbol{\mu}_1 \quad (40)$$

$$\mathbf{F}(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = -\mathbf{f}(\boldsymbol{\theta}_1). \quad (41)$$

Premultiplying equation 40 by $\mathbf{U}^T(\boldsymbol{\theta}_1)$, we have

$$\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = \mathbf{U}^T(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1). \quad (42)$$

Solutions of $(\boldsymbol{\theta} - \boldsymbol{\theta}_1)$ in equation 42 are of the form

$$\boldsymbol{\theta} - \boldsymbol{\theta}_1 = \mathbf{U}(\boldsymbol{\theta}_1)(\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1))^{-1}\mathbf{U}(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1) + \boldsymbol{\eta} \quad (43)$$

where $\boldsymbol{\eta} \in \text{Null}(\mathbf{I}(\boldsymbol{\theta}_1))$. Since, again, $\boldsymbol{\theta}$ is a better approximation of $\hat{\boldsymbol{\theta}}(\mathbf{x})$ than $\boldsymbol{\theta}_1$, this alone motivates an iterative scheme for obtaining the CMLE $\hat{\boldsymbol{\theta}}(\mathbf{x})$. However, if possible, we would like to incorporate equation 41 to eliminate the variable $\boldsymbol{\eta}$ and find the unique solution to the approximated system in equation 39.

Define a continuous map $\mathbf{V} : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times (N-J)}$ such that, for every $\boldsymbol{\theta} \in \mathbb{R}^N$, $\mathbf{V}(\boldsymbol{\theta})$ is a matrix whose columns form an orthonormal basis of the range space of the row vectors of $\mathbf{F}(\boldsymbol{\theta})$. Equivalently, the columns of $\mathbf{V}(\boldsymbol{\theta})$ form an orthonormal basis of the nullspace of $\mathbf{U}(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \mathbb{R}^N$. Thus, $\mathbf{U}^T(\boldsymbol{\theta})\mathbf{V}(\boldsymbol{\theta}) = \mathbf{0}$, $\mathbf{V}^T(\boldsymbol{\theta})\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_{(N-J) \times (N-J)}$ and, consequently, $\mathbf{U}(\boldsymbol{\theta})\mathbf{U}^T(\boldsymbol{\theta}) + \mathbf{V}(\boldsymbol{\theta})\mathbf{V}^T(\boldsymbol{\theta}) = \mathbf{I}_{N \times N}$. Using this identity of $\mathbf{I}_{N \times N}$ in equation 42,

$$\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)(\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1) + \mathbf{V}^T(\boldsymbol{\theta}_1)\mathbf{V}(\boldsymbol{\theta}_1))(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = \mathbf{U}^T(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1). \quad (44)$$

By definition, $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{F}^T(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})$ for some map $\mathbf{R} : \mathbb{R}^N \rightarrow \mathbb{R}^{K \times (N-J)}$, so using equation 41 we have that

$$\mathbf{V}^T(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = \mathbf{R}^T(\boldsymbol{\theta}_1)\mathbf{F}(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = -\mathbf{R}^T(\boldsymbol{\theta}_1)\mathbf{f}(\boldsymbol{\theta}_1).$$

But $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ so $\mathbf{f}(\boldsymbol{\theta}_1) = \mathbf{0}$. Thus, $\mathbf{V}(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = \mathbf{0}$ as well, and equation 44 is now

$$\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1)\mathbf{U}^T(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = \mathbf{U}^T(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1). \quad (45)$$

Multiplying $(\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1))^{-1}$ to both sides of equation 45, we have

$$\mathbf{U}^T(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) = (\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1))^{-1}\mathbf{U}^T(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1).$$

Thus, the additional information from equation 41 eliminates the $\boldsymbol{\eta}$ term in equation 43, and the solution is

$$\begin{aligned} \boldsymbol{\theta} - \boldsymbol{\theta}_1 &= (\mathbf{U}(\boldsymbol{\theta}_1)\mathbf{U}^T(\boldsymbol{\theta}_1) + \mathbf{V}(\boldsymbol{\theta}_1)\mathbf{V}^T(\boldsymbol{\theta}_1))(\boldsymbol{\theta} - \boldsymbol{\theta}_1) \\ &= \mathbf{U}(\boldsymbol{\theta}_1)\mathbf{U}^T(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \boldsymbol{\theta}_1) \\ &= \mathbf{U}(\boldsymbol{\theta}_1)(\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1))^{-1}\mathbf{U}^T(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_1). \end{aligned}$$

So if $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$ is a close approximation of the CMLE $\hat{\boldsymbol{\theta}}(\mathbf{x})$ then $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$ is a closer one. This prompts the following iterative scheme, a variation on the method of scoring

$$\begin{aligned} \boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k + \mathbf{U}(\boldsymbol{\theta}_k)(\mathbf{U}^T(\boldsymbol{\theta}_k)\mathbf{I}(\boldsymbol{\theta}_k)\mathbf{U}(\boldsymbol{\theta}_k))^{-1}\mathbf{U}^T(\boldsymbol{\theta}_k)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k) \\ &= \boldsymbol{\theta}_k + \mathbf{B}(\boldsymbol{\theta}_k)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k) \end{aligned} \tag{46}$$

which, instead of using the CRB as the projection matrix, uses the CCRB⁸. It is important to note here that this is not another Quasi-Newton method, as the corresponding premultiplying matrix \mathbf{D}_k is not positive definite, but rather positive semi-definite. Indeed, for $\mathbf{U}(\boldsymbol{\theta}_k)(\mathbf{U}^T(\boldsymbol{\theta}_k)\mathbf{I}(\boldsymbol{\theta}_k)\mathbf{U}(\boldsymbol{\theta}_k))^{-1}\mathbf{U}^T(\boldsymbol{\theta}_k)$ to be positive definite, $\mathbf{U}(\boldsymbol{\theta}_k)$ would necessarily be full row rank which requires that $\mathbf{F}(\boldsymbol{\theta}_k)$ be null, i.e., an unconstrained model, corresponding to the classical method of scoring.

This constrained scoring formulation, as experienced with the Newton or classical scoring algorithm, does not necessarily lead to convergent sequences. To better control this behavior, an additional variational parameter α_k is introduced to control the step size. The modification is

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{B}(\boldsymbol{\theta}_k)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k) \tag{47}$$

where the step size α_k satisfies some appropriate step size rule that will guarantee usability and will stabilize the convergence. Another motivation for this addition is the lack of knowledge of what Lipschitz condition the objective satisfies. If we did have that knowledge, we could possibly choose a fixed step size $\alpha_k = s$ that enables the iteration to be a local contraction mapping. Since that information is not known, we instead must employ a rule with a variable step size (22), such as:

⁸A special case of equation 46 for linear constraints was presented in (5). And a specific formulation of equation 46 exists for the conventional optimization problem, again with linear constraints, in (20, p. 178). Since that problem is non-statistical, in place of the FIM is the Hessian of the Lagrangian. Our problem is statistical, and so we instead estimate the Hessian with the FIM.

1. A minimization rule.
2. A diminishing step size rule.
3. A successive step-size rule (e.g., the *Armijo rule*).

The minimization rule chooses the optimal α_k that minimizes the corresponding objective for each iterate. This rule relies on a one-line search for each iterate. A possible variation is to restrict α_k to some finite interval. The diminishing step size rule chooses a sequence $\{\alpha_k\}$ with the restriction that $\alpha_k \rightarrow 0$ while $\sum_{k=1}^{\infty} \alpha_k = \infty$. A particular step size, however, might result in a greater negative log-likelihood, so this method still requires a check to guarantee usability. If a particular step is not usable, the rule might be adjusted to skip sufficient elements of the sequence until the step is usable. The Armijo rule chooses a sequence defined by

$$\alpha_k = \beta^{m_k} s \quad (48)$$

where m_k is the first nonnegative integer m such that

$$\log p(\mathbf{x}; \boldsymbol{\theta}_{k+1}) - \log p(\mathbf{x}, \boldsymbol{\theta}_k) \geq \frac{\sigma}{\beta^m s} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k+1}\|^2 \quad (49)$$

and σ , β , and s are fixed scalars with $0 < \sigma < 1$, $0 < \beta < 1$, and $0 < s$. If $\boldsymbol{\theta}_k$ is a stationary point, then α_k is set to 0. The check in equation 49 guarantees that the steps of each iterate are usable.

4.2 The Restoration Step

To eliminate the $\boldsymbol{\eta}$ variable in equation 43, we used the fact that the first iterate $\boldsymbol{\theta}_1$ was a point in the constraint set $\boldsymbol{\Theta}$. However, the next iterate $\boldsymbol{\theta}_2$ is not guaranteed to satisfy the constraints since it is the solution of the Taylor approximations in equations 36 and 37.

The projection matrix is constant, so between iterates the search (over α_k) is one-dimensional or linear, which moves away from any nonlinear constraint. In fact, it is only guaranteed to satisfy the constraints if they are linear. Thus, while the sequence generated by equation 47 may converge, it may not converge to a point in the constraint set $\boldsymbol{\Theta}$. The process to correct this error is called the restoration step.

We restore the second iterate back onto the constraint set $\boldsymbol{\Theta}$ using a projection. Since $\boldsymbol{\Theta}$ is convex, the projection theorem favors using the natural projection for this step. When the projection is the natural one, the projection theorem says it is uniquely determined. Let $\boldsymbol{\pi} : \mathbb{R}^N \rightarrow \boldsymbol{\Theta}$ be the natural projection of \mathbb{R}^N -space onto $\boldsymbol{\Theta}$. Then the method of scoring with constraints, or the CSA, is given by

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k(\alpha_k) = \boldsymbol{\pi}[\boldsymbol{\theta}_k + \alpha_k \mathbf{B}(\boldsymbol{\theta}_k) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k)] \quad (50)$$

where α_k satisfies one of the previously listed step size rules. This is similar to a two-metric projection method which typically has improved performance over simple gradient projections (22).

It is desirable that the projection be a somewhat simple operation, e.g., planar or spherical restrictions, so as not to be a computational burden. However, constraint sets may arise where the projection cannot be expressed analytically. For these scenarios, we present the following scheme to minimize the error away from Θ to a predetermined acceptable level (20).

Suppose $\theta_i^{(1)}$, for some i , is a regular point generated by the RHS of equation 47 not in the constraint set Θ . Then $\mathbf{f}(\theta_i^{(1)}) \neq \mathbf{0}$. Let θ_i be the nearest point to $\theta_i^{(1)}$ that does satisfy the constraints, so that $\mathbf{f}(\theta_i) = \mathbf{0}$. Then a Taylor approximation of $\mathbf{f}(\theta_i) = \mathbf{0}$ about $\theta_i^{(1)}$ evaluated at θ is given by

$$\mathbf{0} = \mathbf{f}(\theta_i^{(1)}) + \mathbf{F}(\theta_i^{(1)})(\theta - \theta_i^{(1)}).$$

As with equation 37, we have in θ a point closer to θ_i than $\theta_i^{(1)}$ is. Solutions of θ are of the form

$$\theta = \theta_i^{(1)} - \mathbf{F}^T(\theta_i^{(1)})(\mathbf{F}(\theta_i^{(1)})\mathbf{F}^T(\theta_i^{(1)}))^{-1}\mathbf{f}(\theta_i^{(1)}) + \zeta$$

where $\zeta \in \text{Null}(\mathbf{U}^T(\theta_i^{(1)}))$. Note the requirement that $\theta_i^{(1)}$ be regular is necessary for the inverse of $\mathbf{F}(\theta_i^{(1)})\mathbf{F}^T(\theta_i^{(1)})$ to exist. The restoration update is then given by

$$\theta_i^{(k+1)} = \theta_i^{(k)} - \mathbf{F}^T(\theta_i^{(k)})(\mathbf{F}(\theta_i^{(k)})\mathbf{F}^T(\theta_i^{(k)}))^{-1}\mathbf{f}(\theta_i^{(k)}) + \zeta. \quad (51)$$

4.3 Implementation of the Constrained Scoring Algorithm

The CSA is in the class of null-space algorithms, exploiting $\mathbf{U}(\theta)$ directly. Instead of employing the Hessian or the FIM as done in Quasi-Newton methods, we employ the singular CCRB matrix to minimize the constrained score function.⁹ In fact, given the unconstrained FIM $\mathbf{I}(\theta)$, equation 46 or 50 provides a simple algorithm to verify CML performance. Otherwise, given a good initialization, equation 50 provides a method to obtain CML performance. This can be considered fine-tuning the estimate that provides the initialization.

If the quadratic negative log-likelihood model holds, and the constraints are linear, then when initialized with a feasible point, the CSA solves the CMLE in a single step. This is shown in detail in section 6.1.

⁹The CCRB is only positive semidefinite, not positive definite, since the null space matrix $\mathbf{U}(\theta)$ can never be full row rank when constraints are applied.

The components of the CSA (see figure 1) include the restoration steps in equation 51 (if the projection is not simple), the matrix inverse component of the CCRB in equation 50, and the evaluation of an expectation (an integration) in the FIM. To reduce the complexity (see table 1) of these tasks we mention some variations of equation 50.

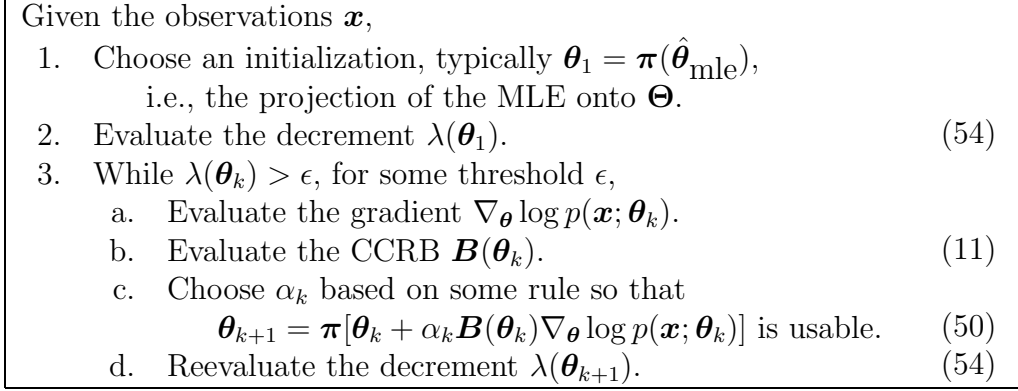


Figure 1. The constrained scoring algorithm.

Table 1. Complexity of the CSA per iteration.

Matrix	Complexity
$\mathbf{U}(\boldsymbol{\theta})$ (given $\mathbf{F}(\boldsymbol{\theta})$)	$< N^6$
$\mathbf{U}^T(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})$	$NJ(N+J)$
$(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}$	J^3
$\mathbf{U}(\boldsymbol{\theta})(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}\mathbf{U}^T(\boldsymbol{\theta})$	$NJ(N+J)$
$\mathbf{B}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta})$	N^2
$\lambda(\boldsymbol{\theta})$	N

If we have a closed form expression for the FIM $\mathbf{I}(\boldsymbol{\theta})$, this eliminates the need to compute an integral for each step. It is, however, highly unlikely to have an exact formulation for the matrix inverse in the CCRB except in trivial or simple cases. A standard procedure in optimization theory for gradient methods is to eliminate the need to compute a matrix inversion for every step by reusing the initial inverse matrix. This adjustment leads to the following variant of the CSA

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k(\alpha_k) = \boldsymbol{\pi}[\boldsymbol{\theta}_k + \alpha_k \mathbf{U}(\boldsymbol{\theta}_k)(\mathbf{U}^T(\boldsymbol{\theta}_1)\mathbf{I}(\boldsymbol{\theta}_1)\mathbf{U}(\boldsymbol{\theta}_1))^{-1}\mathbf{U}^T(\boldsymbol{\theta}_k)\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k)]. \quad (52)$$

In this modified CSA, analogous to the modified Newton's method, the matrix inversion needs to be done only once, for the initial value. This procedure also requires only one evaluation of an expectation (or integral) to obtain the initial FIM. Another standard variation is to apply the inversion after every $j > 1$ iterations.

If the projection is simple or, equivalently, the constraint set simply defined, e.g., planar, spherical, or boundaries, then the restoration step is easily computed. However, if the projection is not easily computed and an iterative scheme as in equation 51 is required, the restoration step is somewhat tedious and may be time-consuming. Alternatively, the CSA might be adjusted by only restoring occasionally. The procedure would then be to obtain via equation 46 the i th iterate and then successively restore the iterate to the constraint set via equation 51, and then to repeat the procedure.

Until now, we have also always assumed that the projection has been onto a convex set. This condition guarantees that each projection in the restoration step is unique; this is shown via the Projection Theorem (18,22). However, note that this condition was not used in section 3, nor was convexity used in the development of the projection step. In either case, only connectivity was required. This is important to point out, since every region defined by a nonlinear equality constraint is nonconvex. However, these nonlinear equality constraints are the ones of practical interest. So if we simply let Θ be a connected set, we can make the following enhancement in the definition of $\pi[\cdot]$ in equation 50. Let $\pi : \mathbb{R}^N \rightarrow \Theta$ be the natural projection of \mathbb{R}^N -space onto Θ with the minimal distance. For practical sets, e.g., a sphere in \mathbb{R}^N , the projection $\pi[\cdot]$ is almost everywhere unique.

Another implementation issue is when to stop the iterations. An obvious choice is to measure the change in the likelihood after each iteration. For Newton's method, another stopping criteria is the *Newton decrement* (18),

$$\lambda(\theta) = (\nabla_{\theta}^T \log p(\mathbf{x}; \theta) (\nabla_{\theta}^2 \log p(\mathbf{x}; \theta))^{-1} \nabla_{\theta} \log p(\mathbf{x}; \theta))^{1/2}. \quad (53)$$

This is a norm with respect to the Hessian of the objective. Note $\lambda(\theta_k)$ decays as the iterations converge to a stationary point. The corresponding decrement for constrained scoring is

$$\lambda(\theta) = (\nabla_{\theta}^T \log p(\mathbf{x}; \theta) \mathbf{B}(\theta) \nabla_{\theta} \log p(\mathbf{x}; \theta))^{1/2}. \quad (54)$$

This scoring decrement is shown to be 0 for stationary points in section 5. By a continuity argument, it follows that the iterations are close to the stationary point when $\lambda(\theta_k)$ is sufficiently small.

In the next section, we show that the CSA at least converges to a local maximum of the likelihood. The only requirement is that $\mathbf{U}^T(\theta) \mathbf{I}(\theta) \mathbf{U}(\theta)$ be positive definite. However, this is also a requirement for the existence of the CCRB (3).

5. Convergence Properties

In this section, we examine convergence properties of the constrained scoring algorithm motivated by the properties for projected gradient presented in (23). The obvious statement regarding convergence is that the sequence of iterates $\{\boldsymbol{\theta}_k\}$ has the CMLE $\hat{\boldsymbol{\theta}}(\mathbf{x})$ as its limit provided the initial guess is sufficiently close. For convenience, we will choose to analyze only the convergence properties of the CSA with the Armijo step size rule, although it is not too difficult to modify these proofs for another rule. By construction, we shall show the algorithm is one that converges to a local stationary point if there is indeed a local minimum.

Let $\{\boldsymbol{\theta}_k\}$ be any sequence generated by the constrained scoring algorithm. Thus, $\boldsymbol{\theta}_1$ is an arbitrarily chosen feasible point, and successive iterates are determined by the CSA in equation 50. (Of course, $\boldsymbol{\theta}_1$ would not be chosen so randomly, but we ignore this aspect of the convergence for the moment.) Define $\Sigma_{\boldsymbol{\theta}_k} = \{\boldsymbol{\theta} \in \Theta | p(\mathbf{x}; \boldsymbol{\theta}) \geq p(\mathbf{x}; \boldsymbol{\theta}_k)\}$.¹⁰

Now, by definition of the chosen Armijo rule equation 49, it can be seen that

$$\log p(\mathbf{x}; \boldsymbol{\theta}_k) \leq \log p(\mathbf{x}; \boldsymbol{\theta}_{k+1})$$

for every positive integer k . Thus, we have the following fact about $\{\boldsymbol{\theta}_k\}$:

Property 1. *The sequence $\{-\log p(\mathbf{x}; \boldsymbol{\theta}_k)\}$ is a monotone decreasing sequence. Furthermore, if $-\log p(\mathbf{x}; \boldsymbol{\theta})$ is bounded below then $\{-\log p(\mathbf{x}; \boldsymbol{\theta}_k)\}$ converges.*

Thus, given an iterate $\boldsymbol{\theta}_k$, all successive iterates are contained in $\Sigma_{\boldsymbol{\theta}_k}$, i.e., $\boldsymbol{\theta}_j \in \Sigma_{\boldsymbol{\theta}_k}$ for all $j \geq k$. The second statement is simply the monotone convergence principle from analysis. And, by the monotonicity of $-\log(\cdot)$, then $\{p(\mathbf{x}; \boldsymbol{\theta}_k)\}$ is also a convergent sequence when $\{-\log p(\mathbf{x}; \boldsymbol{\theta}_k)\}$ is. We will assume that the likelihood is bounded above, so that this sequence always converges. This leads to the following result:

Property 2. *The sequence $\{\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|\}$ vanishes as $k \rightarrow \infty$.*

Proof. Again from equation 49 we have that $\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|^2$ is bounded by the product of an iterate from a sequence bounded below and $\log p(\mathbf{x}; \boldsymbol{\theta}_{k+1}) - \log p(\mathbf{x}; \boldsymbol{\theta}_k)$. Since $\{\log p(\mathbf{x}; \boldsymbol{\theta}_k)\}$ converges, the bound is made arbitrarily small for sufficiently large k . Thus, the same is true for $\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|$. □

¹⁰In this section, we will assume Θ is convex so that $\pi[\cdot]$ is the natural projection.

This does not imply that the sequence $\{\boldsymbol{\theta}_k\}$ converges. However, if $\Sigma_{\boldsymbol{\theta}_k}$ is bounded, then the Bolzano-Weierstrass theorem (19) implies the existence of a cluster, or accumulation, point, i.e., the existence of a subsequence that does converge.

Property 3. *If $\Sigma_{\boldsymbol{\theta}_1}$ is compact, then cluster points of the sequence $\{\boldsymbol{\theta}_k\}$ are also stationary points.*

Recall, a point $\boldsymbol{\theta}$ is stationary if it satisfies the KKT condition equations 20 through 24. The initial iterate is feasible and the restoration $\boldsymbol{\pi}[\cdot]$ preserves feasibility. Since we've incorporated the active constraints into \mathbf{f} , the only additional condition is equation 25. So if $\boldsymbol{\theta}^*$ is a stationary point, then

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*) \mathbf{U}(\boldsymbol{\theta}^*) = \boldsymbol{\mu}^{*T} \mathbf{F}(\boldsymbol{\theta}^*) \mathbf{U}(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Hence, $\mathbf{B}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*) = \mathbf{U}(\boldsymbol{\theta}) (\mathbf{U}^T(\boldsymbol{\theta}) \mathbf{I}(\boldsymbol{\theta}) \mathbf{U}(\boldsymbol{\theta}))^{-1} \mathbf{U}^T(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*) = \mathbf{0}$. Thus, a point is stationary only if

$$\boldsymbol{\theta}^*(\alpha) = \boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha \mathbf{B}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] = \boldsymbol{\pi}(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^*.$$

By definition of the Armijo rule, we also have $\alpha = 0$ at stationary points. While this step is necessary in the method of gradient projection, it is not so here as can be seen above, since the end result would hold regardless.

Proof of Proposition 3. Let $\boldsymbol{\theta}^*$ be a cluster point of the sequence $\boldsymbol{\theta}_k$. Then there exists a convergent subsequence $\boldsymbol{\theta}_{n_k}$ which has $\boldsymbol{\theta}^*$ as its limit. Note, by the triangle inequality,

$$\begin{aligned} & \|\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{U}(\boldsymbol{\theta}^*) (\mathbf{U}^T(\boldsymbol{\theta}^*) \mathbf{I}(\boldsymbol{\theta}^*) \mathbf{U}(\boldsymbol{\theta}^*))^{-1} \mathbf{U}^T(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] - \boldsymbol{\theta}^*\| \\ & \leq \|\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{B}(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] - \boldsymbol{\theta}^*\| \\ & \leq \|\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{B}(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] - \boldsymbol{\pi}[\boldsymbol{\theta}_{n_k} + \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})] \\ & \quad + \boldsymbol{\pi}[\boldsymbol{\theta}_{n_k} + \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})] - \boldsymbol{\theta}_{n_k} + \boldsymbol{\theta}_{n_k} - \boldsymbol{\theta}^*\| \\ & \leq \|\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{B}(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] - \boldsymbol{\pi}[\boldsymbol{\theta}_{n_k} + \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})]\| \\ & \quad + \|\boldsymbol{\pi}[\boldsymbol{\theta}_{n_k} + \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})] - \boldsymbol{\theta}_{n_k}\| + \|\boldsymbol{\theta}_{n_k} - \boldsymbol{\theta}^*\|. \end{aligned} \quad (55)$$

Since the projection $\boldsymbol{\pi}$ is onto a convex set Θ , the distance between two points in \mathbb{R}^N is at least as great as the distance between the projections of those two points. Thus, we have

$$\begin{aligned} & \|\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{B}(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] - \boldsymbol{\pi}[\boldsymbol{\theta}_{n_k} + \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})]\| \\ & \leq \|\boldsymbol{\theta}^* + \alpha^* \mathbf{B}(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*) - \boldsymbol{\theta}_{n_k} - \alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k})\| \end{aligned} \quad (56)$$

Thus, applying the inequality of equation 56 to 55 we have

$$\begin{aligned}
& \|\pi[\theta^* + \alpha^* U(\theta^*)(U^T(\theta^*)I(\theta^*)U(\theta^*))^{-1}U^T(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*)] - \theta^*\| \\
& \leq \|\theta^* + \alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \theta_{n_k} + \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|\pi[\theta_{n_k} + \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})] - \theta_{n_k}\| - \|\theta_{n_k} - \theta^*\| \\
& \leq \|\theta^* - \theta_{n_k}\| + \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|\pi[\theta_{n_k} + \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})] - \theta_{n_k}\| + \|\theta_{n_k} - \theta^*\| \quad (57) \\
& \leq 2\|\theta_{n_k} - \theta^*\| + \|\theta_{n_k+1} - \theta_{n_k}\| \\
& \quad + \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \quad (58)
\end{aligned}$$

where the triangle inequality is applied to obtain equation 57, and the CSA iteration in equation 50 is used to obtain equation 58. Taking the second term in equation 58, and applying the triangle inequality yet again, the last term of equation 58 is bounded as

$$\begin{aligned}
& \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \leq \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k}) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \leq \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k}) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \leq \|\alpha^* B(\theta^*)(\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k}))\| \\
& \quad + \|(\alpha^* B(\theta^*) - \alpha_{n_k} B(\theta_{n_k}))\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\|. \quad (59)
\end{aligned}$$

In Euclidean space (or any normed space), we have the property that $\|M\mathbf{v}\| \leq \|M\| \cdot \|\mathbf{v}\|$ for any matrix M and vector \mathbf{v} (24). Thus equation 59 becomes

$$\begin{aligned}
& \|\alpha^* B(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \alpha_{n_k} B(\theta_{n_k})\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \leq \|\alpha^* B(\theta^*)\| \|\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|(\alpha^* B(\theta^*) - \alpha_{n_k} B(\theta_{n_k}))\| \|\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\|. \quad (60)
\end{aligned}$$

Substituting equation 60 into 58, we have

$$\begin{aligned}
& \|\pi[\theta^* + \alpha^* U(\theta^*)(U^T(\theta^*)I(\theta^*)U(\theta^*))^{-1}U^T(\theta^*)\nabla_{\theta} \log p(\mathbf{x}; \theta^*)] - \theta^*\| \\
& \leq 2\|\theta_{n_k} - \theta^*\| + \|\alpha^* B(\theta^*)\| \|\nabla_{\theta} \log p(\mathbf{x}; \theta^*) - \nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| \\
& \quad + \|(\alpha^* B(\theta^*) - \alpha_{n_k} B(\theta_{n_k}))\| \|\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\| + \|\theta_{n_k+1} - \theta_{n_k}\| \quad (61)
\end{aligned}$$

Since Σ_{θ_1} is compact, then $\|\theta_{n_k}\|$ and $\|\nabla_{\theta} \log p(\mathbf{x}; \theta_{n_k})\|$ are bounded. Note the inequality holds for all n_k , so let $n_k \rightarrow \infty$. Then we have that $\theta_{n_k} \rightarrow \theta^*$, and by continuity,

$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_{n_k}) \rightarrow \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)$ and $\alpha_{n_k} \mathbf{B}(\boldsymbol{\theta}_{n_k}) \rightarrow \alpha^* \mathbf{B}(\boldsymbol{\theta}^*)$. Now the first term of equation 61 vanishes by Property 2. Thus, the right side of equation 61 can be made arbitrarily small and, therefore,

$$\boldsymbol{\pi}[\boldsymbol{\theta}^* + \alpha^* \mathbf{U}(\boldsymbol{\theta}^*)(\mathbf{U}^T(\boldsymbol{\theta}^*) \mathbf{I}(\boldsymbol{\theta}^*) \mathbf{U}(\boldsymbol{\theta}^*))^{-1} \mathbf{U}^T(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)] = \boldsymbol{\theta}^*,$$

i.e., $\boldsymbol{\theta}^*$ is stationary. The statement holds true as well if $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_n}$ is compact for some positive integer n . □

Property 4. *If $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_1}$ is compact for all sequences in a set $\boldsymbol{\Theta}'$ and there is a unique cluster point $\boldsymbol{\theta}_o$ for all such sequences then $\lim_{k \rightarrow \infty} \boldsymbol{\theta}_k = \boldsymbol{\theta}_o$ for every sequence $\{\boldsymbol{\theta}_k\}$. Also, $\boldsymbol{\theta}_o$ is the minimum of $-\log p(\mathbf{x}; \boldsymbol{\theta})$.*

Essentially, if only one accumulation point exists for all such sequences, it must be the CMLE, i.e., $\lim_{k \rightarrow \infty} \boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}(\mathbf{x})$.

6. The Linear Model with Constraints

Suppose we have a vector of observations \mathbf{x} of a parameter vector $\boldsymbol{\vartheta}$ given by the following linear model

$$\mathbf{x} = \mathcal{H}\boldsymbol{\vartheta} + \mathbf{n} \tag{62}$$

where \mathcal{H} is a known observation matrix and \mathbf{n} is random noise from $\mathcal{N}(\mathbf{0}, \mathcal{C})$ with known covariance \mathcal{C} . The true parameter vector will be $\boldsymbol{\vartheta}_o$. To be general, we will assume all parameters are complex-valued. The MLE of this problem is well known (4, pp. 528), and also happens to be the best linear unbiased estimator (BLUE) and the minimum variance unbiased (MVU) estimator:

$$\bar{\boldsymbol{\vartheta}}(\mathbf{x}) = (\mathcal{H}^H \mathcal{C}^{-1} \mathcal{H})^{-1} \mathcal{H}^H \mathcal{C}^{-1} \mathbf{x}. \tag{63}$$

This MLE is also efficient, i.e., it has a covariance matrix equal to the complex CRB

$$\mathcal{C}_{\boldsymbol{\vartheta}} = E_{\boldsymbol{\vartheta}_o}(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_o)(\bar{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_o)^H = (\mathcal{H}^H \mathcal{C}^{-1} \mathcal{H})^{-1} = \mathcal{I}^{-1}(\boldsymbol{\vartheta}_o).$$

where $\mathcal{I}(\boldsymbol{\vartheta})$ is the complex Fisher information matrix (4, p. 529) of the model. Note here that the log-likelihood is quadratic with respect to $\boldsymbol{\vartheta}$; thus the Fisher score is¹¹

$$\nabla_{\boldsymbol{\vartheta}} \log p(\mathbf{x}; \boldsymbol{\vartheta}) = \mathcal{H}^T \mathcal{C}^{-*}(\mathbf{x}^* - \mathcal{H}^* \boldsymbol{\vartheta}^*)$$

¹¹There are numerous definitions of the complex derivative. We define it to be $\frac{\partial}{\partial \alpha} = \frac{1}{2}(\frac{\partial}{\partial(\text{Re}\alpha)} - j\frac{\partial}{\partial(\text{Im}\alpha)})$ for complex-valued α . A benefit of this definition is that numerous results are preserved for strictly real-valued parameters.

and the Hessian is

$$\nabla_{\boldsymbol{\vartheta}}^2 \log p(\mathbf{x}; \boldsymbol{\vartheta}) = -\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{H}} = -\boldsymbol{\mathcal{I}}(\boldsymbol{\vartheta}).$$

Hence, the FIM and CRB are constant in $\boldsymbol{\vartheta}$. Thus, for convenience in this section, we will simply denote the FIM as $\boldsymbol{\mathcal{I}}$.

The complex model can be described in terms of real-valued parameters as well. Define $\boldsymbol{\theta} = [\text{Re}(\boldsymbol{\vartheta})^T, \text{Im}(\boldsymbol{\vartheta})^T]^T$, then the real-valued FIM is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{I} = 2 \begin{bmatrix} \text{Re}(\boldsymbol{\mathcal{I}}) & -\text{Im}(\boldsymbol{\mathcal{I}}) \\ \text{Im}(\boldsymbol{\mathcal{I}}) & \text{Re}(\boldsymbol{\mathcal{I}}) \end{bmatrix} = 2 \begin{bmatrix} \text{Re}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{H}}) & -\text{Im}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{H}}) \\ \text{Im}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{H}}) & \text{Re}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1} \boldsymbol{\mathcal{H}}) \end{bmatrix}.$$

In which case, the Fisher score is now

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) = 2 \begin{bmatrix} \text{Re}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1}(\mathbf{x} - \boldsymbol{\mathcal{H}}\boldsymbol{\vartheta})) \\ \text{Im}(\boldsymbol{\mathcal{H}}^H \boldsymbol{\mathcal{C}}^{-1}(\mathbf{x} - \boldsymbol{\mathcal{H}}\boldsymbol{\vartheta})) \end{bmatrix}.$$

The Hessian, likewise, is still the negative FIM. The necessity of converting the FIM and score to the real parameter case is so we may apply the CSA in the following.

6.1 Linear Constraints

Now, suppose that linear constraints are imposed on the parameters, i.e.,

$$\boldsymbol{\Theta} = \{\boldsymbol{\vartheta} : \boldsymbol{f}(\boldsymbol{\vartheta}) = \boldsymbol{\mathcal{F}}\boldsymbol{\vartheta} + \boldsymbol{\nu} = \mathbf{0}\} \quad (64)$$

where $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\nu}$ are known. We assume that $\boldsymbol{\Theta}$ is nonempty. Then $\boldsymbol{\mathcal{F}}(\boldsymbol{\vartheta}) = \boldsymbol{\mathcal{F}}$ for all $\boldsymbol{\vartheta}$. If we assume that $\boldsymbol{\Theta}$ is regular, i.e., $\boldsymbol{\mathcal{F}}$ is full row rank, and that $\boldsymbol{\mathcal{H}}$ is full column rank, then the CMLE is given by the constrained least squares estimator (CLSE) (4, p. 252)

$$\hat{\boldsymbol{\vartheta}}_{CLSE}(\mathbf{x}) = \bar{\boldsymbol{\vartheta}}(\mathbf{x}) - \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H (\boldsymbol{\mathcal{F}} \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H)^{-1} (\boldsymbol{\mathcal{F}} \bar{\boldsymbol{\vartheta}}(\mathbf{x}) + \boldsymbol{\nu}) \quad (65)$$

$$= \left(\boldsymbol{\mathcal{I}}^{-1} - \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H (\boldsymbol{\mathcal{F}} \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H)^{-1} \boldsymbol{\mathcal{F}} \boldsymbol{\mathcal{I}}^{-1} \right) \boldsymbol{\mathcal{I}} \bar{\boldsymbol{\vartheta}}(\mathbf{x}) - \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H (\boldsymbol{\mathcal{F}} \boldsymbol{\mathcal{I}}^{-1} \boldsymbol{\mathcal{F}}^H)^{-1} \boldsymbol{\nu} \quad (66)$$

The first term of this last equation uses the Marzetta derivation of the CCRB for the expression in the parenthesis, albeit in complex form. Given this knowledge, it can be shown that this estimator is unbiased and efficient, a result that appears to be undocumented in the literature, even for the real parameter case.¹² The second term in

¹²For completeness, a proof of the unbiasedness and efficiency of equation 65 is presented in appendix C.

equation 66 is a specific feasible point of the constraint set. However, this formulation requires both a nonsingular complex FIM \mathcal{I} (or, equivalently, a full column rank \mathcal{H}) and a full row rank \mathcal{F} , conditions which may not hold. Using the CSA to circumvent these conditions essentially turns the problem into a standard nullspace quadratic exercise (13).

First, we need to convert the problem to real-valued parameters. Note that equation 64 is equivalent to the following real-parameter constraint set

$$\Theta' = \left\{ \boldsymbol{\theta} : \mathbf{f}'(\boldsymbol{\theta}) = \mathbf{F}\boldsymbol{\theta} + \mathbf{v} = \mathbf{0}, \text{ where } \mathbf{F} = \begin{bmatrix} \text{Re}(\mathcal{F}) & -\text{Im}(\mathcal{F}) \\ \text{Im}(\mathcal{F}) & \text{Re}(\mathcal{F}) \end{bmatrix}, \mathbf{v} = \begin{bmatrix} \text{Re}(\boldsymbol{\nu}) \\ \text{Im}(\boldsymbol{\nu}) \end{bmatrix} \right\} \quad (67)$$

Let $\mathbf{U}(\boldsymbol{\theta}) = \mathbf{U}$ be the matrix defined by equation 5. It can be shown that if $\mathbf{U}(\boldsymbol{\vartheta}) = \mathbf{U}$ is defined to be the matrix whose columns form an orthonormal basis of \mathcal{F} so that $\mathcal{F}\mathbf{U} = \mathbf{0}$ and $\mathbf{U}^H\mathbf{U} = \mathbf{I}_{J \times J}$, then

$$\mathbf{U} = \begin{bmatrix} \text{Re}(\mathbf{U}) & -\text{Im}(\mathbf{U}) \\ \text{Im}(\mathbf{U}) & \text{Re}(\mathbf{U}) \end{bmatrix}. \quad (68)$$

Let $\boldsymbol{\vartheta}_1$ be any vector satisfying the constraints so that $\boldsymbol{\vartheta}_1 \in \Theta$. For example, if the space is still regular (\mathcal{F} is full row rank), then $\boldsymbol{\vartheta}_1 = -\mathcal{F}^H(\mathcal{F}\mathcal{F}^H)^{-1}\boldsymbol{\nu}$ is such a point vector, otherwise let $\boldsymbol{\vartheta}_1 = -\mathcal{F}^\dagger\boldsymbol{\nu}$; however, note that any feasible point vector will work. Note, $\boldsymbol{\theta}_1 = [\text{Re}(\boldsymbol{\vartheta}_1)^T, \text{Im}(\boldsymbol{\vartheta}_1)^T]^T$ is the corresponding point in Θ' . As stated earlier, since the log-likelihood is quadratic and the constraints linear, the equations 40 and 41 are exact, so the restoration step and the variable step size in the CSA equation 50 are unnecessary, i.e., the CMLE is found in one step to be the linear estimator given by

$$\begin{aligned} \hat{\boldsymbol{\theta}}(x) &= \boldsymbol{\theta}_1 + \mathbf{B}(\boldsymbol{\theta}_1)\nabla_{\boldsymbol{\theta}} \log p(x; \boldsymbol{\theta}_1) \\ &= \boldsymbol{\theta}_1 + \mathbf{U}(\mathbf{U}^T \mathbf{I}(\boldsymbol{\theta}_1) \mathbf{U})^{-1} \mathbf{U}^T \nabla_{\boldsymbol{\theta}} \log p(x; \boldsymbol{\theta}_1) \\ &= \begin{bmatrix} \text{Re}(\boldsymbol{\vartheta}_1) \\ \text{Im}(\boldsymbol{\vartheta}_1) \end{bmatrix} + \begin{bmatrix} \text{Re}(\mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\boldsymbol{\vartheta}_1)) \\ \text{Im}(\mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\boldsymbol{\vartheta}_1)) \end{bmatrix} \\ \Rightarrow \hat{\boldsymbol{\vartheta}}(x) &= \boldsymbol{\vartheta}_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\boldsymbol{\vartheta}_1) \end{aligned}$$

where $\mathcal{B} = \mathcal{B}(\boldsymbol{\vartheta}) = \mathbf{U}(\mathbf{U}^T \mathcal{I}(\boldsymbol{\vartheta}) \mathbf{U})^{-1} \mathbf{U}^T$ is defined similarly as in equation 11 to be the *complex CCRB* for this constrained problem.¹³ Comparing this CMLE result with the prior CLSE result in equation 65, we see the only requirement now is that $\mathcal{H}\mathbf{U}$ be full column rank, whereas equation 65 required both \mathcal{H} to be full column rank and \mathcal{F} to be full row rank. In other words, the prior solution requires information-regularity and a regular point solution in the original problem; the solution given here only requires the alternative information-regularity condition (see Theorem 1 or (3)). This leads to the following result.

¹³It should be emphasized that the complex CCRB is of the form given by \mathcal{B} only for this particular choice of constraint. For general constraints, the covariance matrix of the real-parameter estimator may not assume the necessary form (4, pp. 524-532).

Theorem 3. *If the observations \mathbf{x} are described by a linear model of the form*

$$\mathbf{x} = \mathbf{H}\boldsymbol{\vartheta} + \mathbf{n}$$

where \mathbf{H} is a known matrix, $\boldsymbol{\vartheta}$ is an unknown parameter subject to the linear constraint $\mathbf{f}(\boldsymbol{\vartheta}) = \mathbf{F}\boldsymbol{\vartheta} + \boldsymbol{\nu} = \mathbf{0}$ with the true parameter being $\boldsymbol{\vartheta}_o$, and \mathbf{n} is a noise vector with PDF $\mathcal{N}(\mathbf{0}, \mathbf{C})$, then provided $\mathbf{H}\mathbf{U}$ is full column rank where \mathbf{U} is given by equation 5, the CMLE of $\boldsymbol{\vartheta}_o$ is given by

$$\hat{\boldsymbol{\vartheta}}(\mathbf{x}) = \boldsymbol{\vartheta}_1 + \mathbf{B}\mathbf{H}^H\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}_1) \quad (69)$$

where \mathbf{B} is the CCRB and $\boldsymbol{\vartheta}_1$ is any arbitrary feasible point (e.g., $\boldsymbol{\vartheta}_1 = \mathbf{F}^\dagger \boldsymbol{\nu}$). $\hat{\boldsymbol{\vartheta}}(\mathbf{x})$ is unbiased and is an efficient estimator which attains the CCRB and, therefore, is the BLUE and the MVU estimator. Furthermore, when \mathbf{H} is full column rank and \mathbf{F} is full row rank, then $\hat{\boldsymbol{\vartheta}}(\mathbf{x}) \equiv \hat{\boldsymbol{\vartheta}}_{CLSE}(\mathbf{x})$.

Proof. First note that since $\mathbf{F}\mathbf{B} = \mathbf{0}$, the estimator satisfies the constraints:

$$\mathbf{f}(\hat{\boldsymbol{\vartheta}}(\mathbf{x})) = \mathbf{F}(\boldsymbol{\vartheta}_1 + \mathbf{B}\mathbf{H}^H\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}_1)) + \boldsymbol{\nu} = \mathbf{F}\boldsymbol{\vartheta}_1 + \boldsymbol{\nu} = \mathbf{f}(\boldsymbol{\vartheta}_1) = \mathbf{0}.$$

Next, from equation 41 $\boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1 = \mathbf{U}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \mathbb{C}^J$ since by a Taylor expansion $\mathbf{0} = \mathbf{f}(\boldsymbol{\vartheta}_o) = \mathbf{F} \cdot (\boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1)$. Hence, we have the following convenient property:

$$\begin{aligned} \mathbf{B}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}(\boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1) &= \mathbf{U}(\mathbf{U}^H\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H}\mathbf{U})^{-1}\mathbf{U}^H\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H}\mathbf{U}\boldsymbol{\eta} \\ &= \mathbf{U}\boldsymbol{\eta} \\ &= \boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1. \end{aligned}$$

Thus, the CMLE is also unbiased:

$$\begin{aligned} E_{\boldsymbol{\vartheta}_o} \hat{\boldsymbol{\vartheta}}(\mathbf{x}) &= \boldsymbol{\vartheta}_1 + \mathbf{B}\mathbf{H}^H\mathbf{C}^{-1}E_{\boldsymbol{\vartheta}_o}(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}_1) \\ &= \boldsymbol{\vartheta}_1 + \mathbf{B}\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H}(\boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1) \\ &= \boldsymbol{\vartheta}_1 + \boldsymbol{\vartheta}_o - \boldsymbol{\vartheta}_1 \\ &= \boldsymbol{\vartheta}_o \end{aligned}$$

and the estimator is efficient, i.e., its covariance matrix is the CCRB:

$$\begin{aligned}
& E_{\vartheta_o}(\hat{\vartheta}(x) - E_{\vartheta_o}\hat{\vartheta}(x))(\hat{\vartheta}(x) - E_{\vartheta_o}\hat{\vartheta}(x))^H \\
&= E_{\vartheta_o}\hat{\vartheta}(x)\hat{\vartheta}^H(x) - \vartheta_o\vartheta_o^H \\
&= E_{\vartheta_o}[(\vartheta_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\vartheta_1))(\vartheta_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\vartheta_1))^H] - \vartheta_o\vartheta_o^H \\
&= E_{\vartheta_o}[(\vartheta_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(n + \mathcal{H}(\vartheta_o - \vartheta_1))(\vartheta_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(n + \mathcal{H}(\vartheta_o - \vartheta_1)))^H] - \vartheta_o\vartheta_o^H \\
&= \vartheta_1\vartheta_1^H + \vartheta_1(\vartheta_o - \vartheta_1)^H\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}\mathcal{B} + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}E_{\vartheta_o}nn^H\mathcal{C}^{-1}\mathcal{H}\mathcal{B} \\
&\quad + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}(\vartheta_o - \vartheta_1)\vartheta_1^H + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}(\vartheta_o - \vartheta_1)(\vartheta_o - \vartheta_1)^H\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}\mathcal{B} - \vartheta_o\vartheta_o^H \\
&= \vartheta_1\vartheta_1^H + \vartheta_1(\vartheta_o - \vartheta_1)^H + (\vartheta_o - \vartheta_1)\vartheta_1^H + (\vartheta_o - \vartheta_1)(\vartheta_o - \vartheta_1)^H - \vartheta_o\vartheta_o^H \\
&\quad + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}\mathcal{C}\mathcal{C}^{-1}\mathcal{H}\mathcal{B} \\
&= \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}\mathcal{B} \\
&= \mathcal{U}(\mathcal{U}^H\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H}\mathcal{U})^{-1}\mathcal{U}^H = \mathcal{U}(\mathcal{U}^H\mathcal{I}\mathcal{U})^{-1}\mathcal{U}^H = \mathcal{B}.
\end{aligned}$$

So this more general CMLE is also the BLUE as well as the MVU estimator in the general linear model under the linear constraint. To see equivalence to the CLSE expression equation 65, assume that the FIM \mathcal{I} is nonsingular and the constraint space Θ is regular. Then, $\mathcal{B} = \mathcal{I}^{-1} - \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}\mathcal{F}\mathcal{I}^{-1}$, i.e., the Stoica CCRB formulation and the Marzetta formulation are identical (3), and so

$$\begin{aligned}
\hat{\vartheta}(x) &= \vartheta_1 + \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}(x - \mathcal{H}\vartheta_1) \\
&= \mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}x - (\mathcal{B}\mathcal{H}^H\mathcal{C}^{-1}\mathcal{H} - \mathcal{I}_{N \times N})\vartheta_1 \\
&= \mathcal{B}\mathcal{I}\bar{\vartheta}(x) - (\mathcal{B}\mathcal{I} - \mathcal{I}_{N \times N})\vartheta_1 \\
&= \mathcal{B}\mathcal{I}\bar{\vartheta}(x) + \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}\mathcal{F}\vartheta_1 \\
&= \mathcal{B}\mathcal{I}\bar{\vartheta}(x) - \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}\nu \\
&= (\mathcal{I}^{-1} - \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}\mathcal{F}\mathcal{I}^{-1})\mathcal{I}\bar{\vartheta}(x) - \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}\nu \\
&= \bar{\vartheta}(x) - \mathcal{I}^{-1}\mathcal{F}^H(\mathcal{F}\mathcal{I}^{-1}\mathcal{F}^H)^{-1}(\mathcal{F}\bar{\vartheta}(x) + \nu) \\
&= \hat{\vartheta}_{CLSE}(x).
\end{aligned}$$

□

In addition to being an alternative CMLE, equation 69 is also an alternative CLSE.

6.2 Nonlinear Constraints

As an example of nonlinear parametric constraints imposed on equation 62, we consider the constraint set $\Theta = \{\theta : f_i(\vartheta_i) = |\vartheta_i|^2 - 1 = 0, i = 1, 2, \dots, N\}$ where all the elements of ϑ are restricted to be of unit modulus. This was one of the constraints applied in (12) in the evaluation of performance bounds of a different model. It should be noted that this set is

not convex, but natural projections from \mathbb{R}^{2N} onto Θ are a.e. unique, i.e., unique except on a set $\{\boldsymbol{\theta} : \vartheta_i = 0 \text{ for any } i\}$ of measure zero. The gradient matrix of the constraints is then

$$\mathbf{F}(\boldsymbol{\theta}) = 2 \begin{bmatrix} \text{Re}(\mathbf{T}(\boldsymbol{\vartheta})) & \text{Im}(\mathbf{T}(\boldsymbol{\vartheta})) \end{bmatrix}$$

where $\mathbf{T}(\boldsymbol{\vartheta}) = \text{diag}(\boldsymbol{\vartheta})$, i.e., a diagonal matrix with i th row-column element ϑ_i . A matrix whose columns form an orthonormal null space of $\mathbf{F}(\boldsymbol{\theta})$ was found in (12) to be

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{bmatrix} \text{Im}(\mathbf{T}(\boldsymbol{\vartheta})) \\ -\text{Re}(\mathbf{T}(\boldsymbol{\vartheta})) \end{bmatrix}.$$

This results in the following CCRB:

$$\begin{aligned} \mathbf{B}(\boldsymbol{\theta}) &= \mathbf{U}(\boldsymbol{\theta})(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}\mathbf{U}^T(\boldsymbol{\theta}) \\ &= \frac{1}{2} \begin{bmatrix} \text{Im}(\mathbf{T}(\boldsymbol{\vartheta})) \\ -\text{Re}(\mathbf{T}(\boldsymbol{\vartheta})) \end{bmatrix} \cdot (\text{Re}(\mathbf{T}^H(\boldsymbol{\vartheta})\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H}\mathbf{T}(\boldsymbol{\vartheta})))^{-1} \cdot \begin{bmatrix} \text{Im}(\mathbf{T}(\boldsymbol{\vartheta})) & -\text{Re}(\mathbf{T}(\boldsymbol{\vartheta})) \end{bmatrix}. \end{aligned}$$

Thus the CSA is given by

$$\begin{aligned} \boldsymbol{\theta}_{k+1} &= \boldsymbol{\pi} [\boldsymbol{\theta}_k + \alpha_k \mathbf{B}(\boldsymbol{\theta}_k) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k)] \\ &= \boldsymbol{\pi} \left[\boldsymbol{\theta}_k + \alpha_k \begin{bmatrix} \text{Im}(\mathbf{T}(\boldsymbol{\vartheta}_k)) \\ -\text{Re}(\mathbf{T}(\boldsymbol{\vartheta}_k)) \end{bmatrix} \times \right. \\ &\quad \left. (\text{Re}(\mathbf{T}^H(\boldsymbol{\vartheta}_k)\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H}\mathbf{T}(\boldsymbol{\vartheta}_k)))^{-1} \cdot \text{Im}(\mathbf{T}^H(\boldsymbol{\vartheta}_k)\mathbf{H}^H\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}_k)) \right] \end{aligned}$$

where $\boldsymbol{\pi}$ is the projection onto Θ and the relation $\boldsymbol{\theta}_k = [\text{Re}(\boldsymbol{\vartheta}_k)^T, \text{Im}(\boldsymbol{\vartheta}_k)^T]^T$ holds for all k .

For simulation, we randomly selected the complex elements for an 8×8 observation matrix \mathbf{H} from the standard normally distributed number generator provided in MATLAB, and randomly generated unit modulus elements for the constrained $\boldsymbol{\theta}$ vector. We consider the average performance of the \mathbf{C}^{-1} -norm of $(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta})$ and average performance of the mean-square error (MSE) of $\boldsymbol{\vartheta}$ over $n = 5000$ realizations of the noise \mathbf{n} modeled as spatially white; so $\mathbf{C} = \sigma^2 \mathbf{I}_{8 \times 8}$ with $\sigma^2 = \frac{1}{10}$. In the CSA we employ the diminishing step size rule with $\alpha_k = \frac{1}{k}$. (All steps were usable.) For comparison we temporarily ignored the stopping criteria by choosing an infinitesimal bound requirement on the decrement. A reasonable, close initial estimate of the CMLE is the projection of the MLE equation 63 onto Θ , i.e.,

$$\boldsymbol{\vartheta}_1 = \boldsymbol{\pi}[\boldsymbol{\vartheta}(\mathbf{x})] = \boldsymbol{\pi} \left[(\mathbf{H}^H\mathbf{C}^{-1}\mathbf{H})^{-1} \mathbf{H}^H\mathbf{C}^{-1}\mathbf{x} \right].$$

This initial value, as one would expect, turns out in general not to be the CMLE. However, the estimate is often (but not always) a very good initialization, as revealed in figure 2. The norm error, given by $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{H}\boldsymbol{\vartheta}\|$, vs. the iterate is nearly identical for any

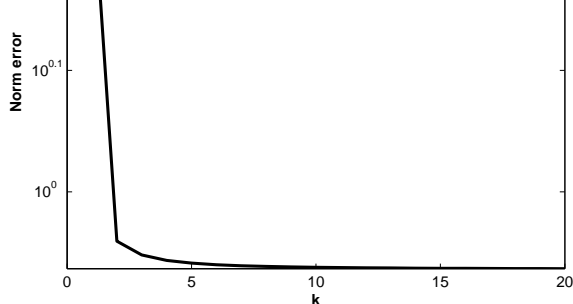


Figure 2. The average norm of $\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}_k$ at iteration k . There is significant gain in the first iterate compared with later iterates, as expected with a quadratically convergent sequence.

particular realization. This metric is key, since seeking to minimize the negative log-likelihood of the linear model equation 62 is equivalent to minimizing the norm of $(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta})$. So we see that the CSA does indeed maximize, at least locally, the log-likelihood. A random local search could not provide a better global maximum, but does make evident the need for a sufficiently close initialization to the CMLE for the CSA (or any Newton-type algorithm) to successfully work in this example. As is evident, since the negative log-likelihood is quadratic and the constraints nearly linear (in a neighborhood of the CMLE) there is significant gain after merely the first one or two iterations as convergence is nearly quadratic. Figure 3 compares the average MSE (per real parameter coefficient) at iteration k to the CCRB given by $\frac{1}{16n} \sum_{i=1}^n \left\| \hat{\boldsymbol{\vartheta}}_k(\mathbf{x}_i) - \boldsymbol{\vartheta}_o \right\|^2$; as can be seen, the CMLE for this scenario achieves the CCRB. The numerical simulations also show this particular CMLE to be asymptotically unbiased in figure 4. We calculated the average of the absolute bias of the parameters, i.e.,

$$\text{average bias} = \frac{1}{16} \sum_{j=1}^{16} \left| \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\theta}}(\mathbf{x}_i)_j - \boldsymbol{\theta}_j \right|,$$

where $\hat{\boldsymbol{\theta}}(\mathbf{x}_i)_j$ and $\boldsymbol{\theta}_j$ refer to the j th element of the vector (and not the iteration in this case). An average error of .005 in both axes of the Cartesian plane roughly corresponds to an average of less than a half-degree of error in the phases of each element of $\boldsymbol{\vartheta}$.

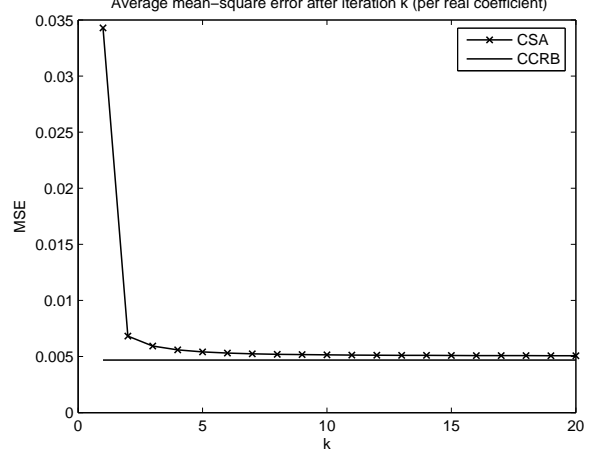
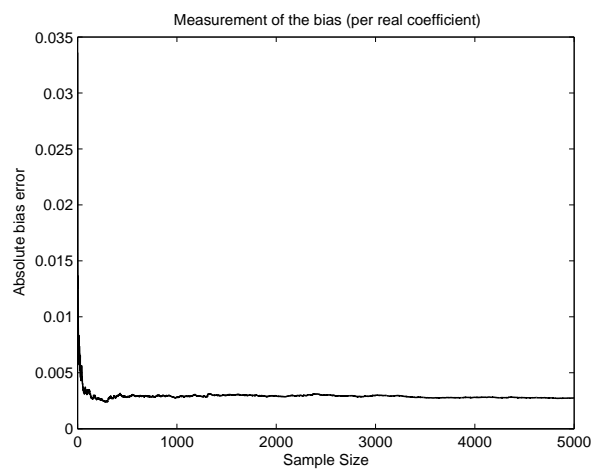


Figure 3. Average mean-s



RB.

Figure 4. Average bias error of the CSA.

7. A Nonlinear Model Example

By a nonlinear model, we mean a model which is nonlinear with respect to the parameter vector $\boldsymbol{\vartheta}$ or $\boldsymbol{\theta}$. Unlike in the previous section, there is no inherent guarantee that maximizing the likelihood also minimizes the MSE of the parameter vector. It depends on the model. However, we still have the property that if an efficient estimator exists for the constrained problem then it must be the CMLE. Recall the condition for equality with the CCRB in Theorem 1:

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta} = \mathbf{B}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}). \quad (70)$$

This implies that

$$\mathbf{0} = \hat{\boldsymbol{\theta}}(\mathbf{x}) - \hat{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{B}(\hat{\boldsymbol{\theta}}(\mathbf{x})) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x})). \quad (71)$$

When $\mathbf{U}^T(\hat{\boldsymbol{\theta}}(\mathbf{x})) \mathbf{I}(\hat{\boldsymbol{\theta}}(\mathbf{x})) \mathbf{U}(\hat{\boldsymbol{\theta}}(\mathbf{x}))$ is positive semidefinite, this implies that $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x})) \in \text{span}(\mathbf{F}^T(\hat{\boldsymbol{\theta}}(\mathbf{x})))$, which satisfies the Lagrangian equation 34. Thus, as with the ML method, the CML method also produces an efficient estimator if it exists.

In this section we examine a scenario given in (12).

7.1 MIMO Instantaneous Mixing Model

Consider the multi-input, multi-output (MIMO) instantaneous mixing model where \mathbf{x}_n is a vector of observations of a linear mixing of unknown parameters at time $n = 1, \dots, N$ given by the model

$$\mathbf{x}_n = \boldsymbol{\mathcal{H}} \mathbf{s}_n + \mathbf{n} \quad (72)$$

where $\boldsymbol{\mathcal{H}}$ is an unknown $M \times K$ complex-valued channel matrix, each \mathbf{s}_n is a complex-valued data symbol vector, and the additive noise vector \mathbf{n} is spatially and temporally white with variance σ^2 . We define a vector of unknown parameters by

$$\boldsymbol{\vartheta} = \begin{bmatrix} \mathbf{h}^{(1)} \\ \mathbf{s}^{(1)} \\ \vdots \\ \mathbf{h}^{(K)} \\ \mathbf{s}^{(K)} \end{bmatrix},$$

where $\mathbf{h}^{(k)}$ is the k th column of $\boldsymbol{\mathcal{H}}$ and $\mathbf{s}^{(k)} = [s_1(k), \dots, s_N(k)]$. Then the complex FIM was found in (25) to be

$$\mathcal{I}(\boldsymbol{\vartheta}) = \frac{2}{\sigma^2} \boldsymbol{\mathcal{Q}}^H \boldsymbol{\mathcal{Q}}$$

where $\mathbf{Q} = [\mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(K)}]$ with $\mathbf{Q}^{(k)} = [\mathbf{I}_{M \times M} \otimes \mathbf{s}^{(k)}, \mathbf{h}^{(k)} \otimes \mathbf{I}_{M \times M}]$. As detailed in (25), this FIM is singular due to the multiplicative ambiguity inherent in the model. As before, the model can be described in terms of the real-valued parameter vector $\boldsymbol{\theta} = [\text{Re}(\boldsymbol{\vartheta})^T \text{Im}(\boldsymbol{\vartheta})^T]^T$, which has a real-valued FIM given by

$$\mathbf{I}(\boldsymbol{\theta}) = 2 \begin{bmatrix} \text{Re}(\mathbf{J}(\boldsymbol{\vartheta})) & -\text{Im}(\mathbf{J}(\boldsymbol{\vartheta})) \\ \text{Im}(\mathbf{J}(\boldsymbol{\vartheta})) & \text{Re}(\mathbf{J}(\boldsymbol{\vartheta})) \end{bmatrix} = \frac{4}{\sigma^2} \begin{bmatrix} \text{Re}(\mathbf{Q}^H \mathbf{Q}) & -\text{Im}(\mathbf{Q}^H \mathbf{Q}) \\ \text{Im}(\mathbf{Q}^H \mathbf{Q}) & \text{Re}(\mathbf{Q}^H \mathbf{Q}) \end{bmatrix}.$$

By the structure of this matrix nullity($\mathbf{I}(\boldsymbol{\theta})$) = 2 · nullity($\mathbf{J}(\boldsymbol{\vartheta})$), so information-regularity cannot be achieved without constraints. Sadler, et al., applied constant modulus and semi-blind constraints on equation 72 to obtain a locally identifiable model (12). The constraint set is

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} : f_{k,n}(\boldsymbol{\theta}) = |s_n(k)|^2 - 1 = 0, \forall k, n; f_t(\boldsymbol{\theta}) = s_t(k) - s_{t,k} = 0, \forall k, t = 1, \dots, T\}$$

where the $s_{t,k}$ are known. For this constraint set, the gradient matrix of the constraints was found in (7) to be

$$\mathbf{F}(\boldsymbol{\theta}) = [\text{Re}(\mathcal{F}(\boldsymbol{\vartheta})) \quad \text{Im}(\mathcal{F}(\boldsymbol{\vartheta}))]$$

where

$$\mathcal{F}(\boldsymbol{\vartheta}) = \begin{bmatrix} \mathcal{F}^{(1)}(\mathbf{s}^{(1)}, \mathbf{h}^{(1)}) & & \\ & \ddots & \\ & & \mathcal{F}^{(K)}(\mathbf{s}^{(K)}, \mathbf{h}^{(K)}) \end{bmatrix} \text{ and } \mathcal{F}^{(k)}(\mathbf{s}^{(k)}, \mathbf{h}^{(k)}) = 2 \begin{bmatrix} \mathbf{I}_{T \times T} & \mathbf{0} & \mathbf{0} \\ j\mathbf{I}_{T \times T} & & \\ \mathcal{D}(\mathbf{s}^{(1)}) & & \mathbf{0} \end{bmatrix},$$

with $\mathcal{D}(\mathbf{s}^{(1)})$ defined as in section 6.2. (Note that this construction of the gradient matrix $\mathbf{F}(\boldsymbol{\theta})$ is not regular since $\boldsymbol{\Theta}$ contains redundancy in restricting some signal values to be unit modulus as well as known values. $\boldsymbol{\Theta}$ might be reformulated so the points are regular, but this is unnecessary.) A matrix whose columns form an orthonormal null space of $\mathbf{F}(\boldsymbol{\theta})$ is given by

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{bmatrix} \text{Im}(\mathcal{U}(\boldsymbol{\vartheta})) \\ -\text{Re}(\mathcal{U}(\boldsymbol{\vartheta})) \end{bmatrix}$$

where

$$\mathcal{U}(\boldsymbol{\vartheta}) = \begin{bmatrix} \mathcal{U}^{(1)}(\mathbf{s}^{(1)}, \mathbf{h}^{(1)}) & & \\ & \ddots & \\ & & \mathcal{U}^{(K)}(\mathbf{s}^{(K)}, \mathbf{h}^{(K)}) \end{bmatrix}$$

and

$$\mathcal{U}^{(k)}(\mathbf{s}^{(k)}, \mathbf{h}^{(k)}) = \begin{bmatrix} \mathbf{0} & -\mathbf{I}_{M \times M} & j\mathbf{I}_{M \times M} \\ \mathcal{D}_T(\mathbf{s}^{(k)}) & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathcal{D}_T(\mathbf{s}^{(k)})$ is the matrix $\mathcal{D}(\mathbf{s}^{(k)})$ with the first T rows removed. This results in the following CCRB:

$$\begin{aligned} \mathbf{B}(\boldsymbol{\theta}) &= \mathbf{U}(\boldsymbol{\theta})(\mathbf{U}^T(\boldsymbol{\theta})\mathbf{I}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta}))^{-1}\mathbf{U}^T(\boldsymbol{\theta}) \\ &= \frac{\sigma^2}{4} \begin{bmatrix} \text{Im}(\mathbf{U}(\boldsymbol{\vartheta})) \\ -\text{Re}(\mathbf{U}(\boldsymbol{\vartheta})) \end{bmatrix} \cdot (\text{Re}(\mathbf{U}^H(\boldsymbol{\vartheta})\mathbf{Q}\mathbf{Q}^H\mathbf{U}(\boldsymbol{\vartheta})))^{-1} \cdot [\text{Im}(\mathbf{U}(\boldsymbol{\vartheta})) \quad -\text{Re}(\mathbf{U}(\boldsymbol{\vartheta}))]. \end{aligned}$$

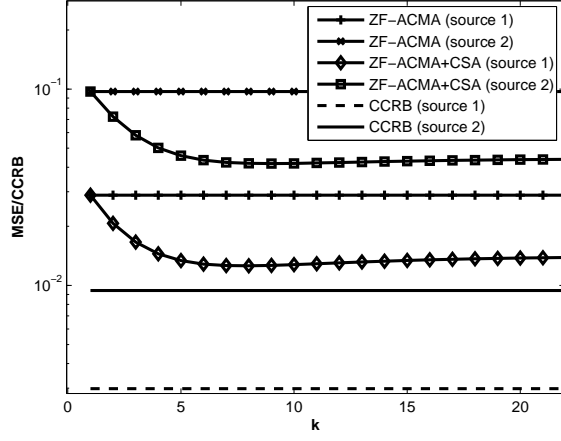
Note the similarity in the structure of the CCRB with that in section 6.2. This is, generally, the form of the CCRB when the appropriate $\mathbf{U}(\boldsymbol{\vartheta})$ matrix can be found. Note that here $\mathbf{U}(\boldsymbol{\vartheta})$ is not the null space of $\mathcal{F}(\boldsymbol{\vartheta})$, but rather these matrices are convenient expressions to formulate $\mathbf{U}(\boldsymbol{\theta})$ and $\mathbf{F}(\boldsymbol{\theta})$, respectively.

We simulated a $K = 2$ source, $M = 2$ channel model over $N = 30$ time samples. The channel \mathcal{H} is taken from the three-ray multipath case in (12) with directions-of-arrival (DOAs) $\{-1, 0, 4\}$ and corresponding complex amplitudes $\{\sqrt{0.2}\angle\frac{-\pi}{6}, \sqrt{0.5}, \sqrt{0.15}\angle\frac{-\pi}{5}\}$ for source 1 and DOAs $\{0, 5, 11\}$ and amplitudes $\{\sqrt{0.15}\angle\frac{-\pi}{5}, \sqrt{0.6}, \sqrt{0.25}\angle\frac{\pi}{3}\}$ for source 2. The source elements were taken randomly from an 8-PSK alphabet. The constraints are the modulus constraint as well as knowledge of the first $T = 2$ symbols for each source discussed above. (FIM-regularity requires at least $T = 1$ training samples per source.) Source 1 (i.e., $\mathbf{s}^{(1)}$) is normalized to have an SNR of 20 dB and the SNR of source 2 is set at 15 dB. We can scale the channel to reflect these signal powers, i.e., for $i = 1, 2$

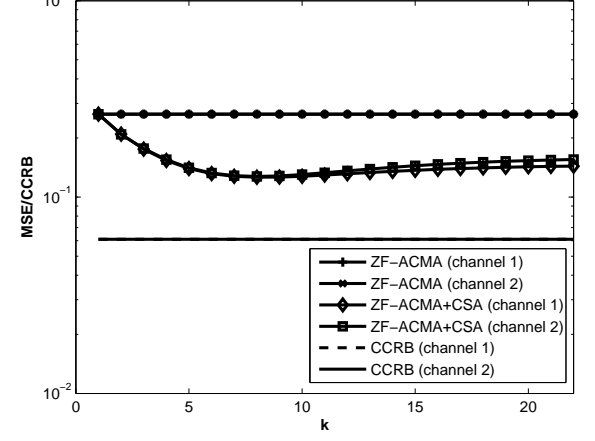
$$\text{SNR}_i = \frac{\|\mathbf{h}^{(i)}\|^2}{M\sigma^2},$$

which allows us to set the noise covariance to be $\sigma^2 = 1$. We obtain an initialization via the zero-forcing variant of the analytical constant modulus algorithm (ZF-ACMA) found in (26). This estimate is projected onto the constraint set $\boldsymbol{\Theta}$ and then we apply the CSA using a successive step size scheme ($\alpha_k = \frac{1}{2^m}$ for the smallest positive integer m that results in a usable step). We calculate the average MSE (per real coefficient) at each iteration over $n = 2000$ trials and compare with the mean CRB for each of the sources and channels.

The numerical simulations show the improvement in the average MSE of both the signals and the channels (figure 5), on average halving the MSE of the initialization estimates provided by ACMA in this instance. Note that after a certain number of iterations, the MSE for both the channel and source elements are actually slightly increasing. This occurs when the decrement $\lambda(\boldsymbol{\theta}_k)$ is sufficiently close to 0 (see figure 6). Setting the stopping criteria to be, e.g., $\lambda(\boldsymbol{\theta}_k) < 1$, achieves the least MSE from the CSA iterations. We note that generally there is greater improvement for the channel estimates over the signal estimates. Also, we note that the CSA does not appear to reduce the MSE significantly for the $T = 1$ case or for low SNR values.



(a)



(b)

Figure 5. Average MSE of the elements compared with the mean.

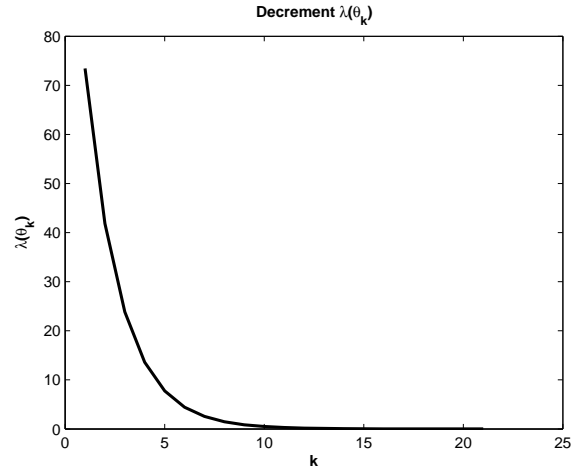


Figure 6. The average decrement at each iteration.

8. Conclusions

We determined the asymptotic normality properties of a parametrically constrained maximum-likelihood problem. In doing so, we have shown that this CMLE is consistent and asymptotically efficient. We then derived a generalization of the method of scoring, the Constrained Scoring Algorithm, using Lagrangian optimization methods. The CSA maintains certain desirable convergence properties and proofs of these have been offered. Finally, we examined several problems of interest: the classical linear model and an instantaneous linear mixing model, to verify the usefulness of this approach. For the linear model, we found another CMLE which we have shown to be unbiased and efficient.

References

- [1] Gorman, J. D.; Hero, A. O. Lower bounds for parametric estimation with constraints. *IEEE Transactions on Information Theory* **1990**, *26* (6), 1285–1301.
- [2] Marzetta, T. L. A simple derivation of the constrained multiple parameter Cramer-Rao bound. *IEEE Transactions on Signal Processing* **1993**, *41* (6), 2247–2249.
- [3] Stoica, P.; Ng, B. C. On the Cramér-Rao Bound under parametric constraints. **1998** *5* (7).
- [4] Kay, S. M. *Fundamentals of Statistical Signal Processing, Estimation Theory*, Prentice-Hall, 1993.
- [5] Osborne, Michael R. Scoring with constraints. *ANZIAM Journal* **2000** *42* (1), 9–25.
- [6] Osborne, Michael R. Fisher’s method of scoring. *Int. Stat. Rev.* **1992** *60* 99–117.
- [7] Kozick, Richard J.; Sadler, Brian M.; Moore, Terrence J. Performance of MIMO: CM and semi-Blind cases *2003 4th IEEE Workshop on Signal Processing Advances in Wireless Communications*, **2003**, 309–13.
- [8] Leshem, Amir. Maximum likelihood separation of constant modulus signals. *IEEE Transactions on Signal Processing*, **2000**, *48* (10), 2948–52.
- [9] Atchinson, J.; Silvey, S. D. Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **1958**, *29*, 813–828.
- [10] Atchinson, J.; Silvey, S. D. Maximum-likelihood estimation procedures and associated tests of significance *J. R. Statist. Soc. B*, **1960** *22*, 154–71.
- [11] Kay, S. M. *Fundamentals of Statistical Signal Processing, Detection Theory*, Prentice-Hall, 1998.
- [12] Sadler, Brian M.; Kozick, Richard J.; Moore, Terrence J. *On the performance of source separation with constant modulus signals*; ARL-TR-3462; U.S. Army Research Laboratory: Adelphi, MD, March 2005.
- [13] Gill, Philip E.; Murray, Walter; Wright, Margaret H. *Practical Optimization*, Academic Press, 1981.
- [14] Luenberger, David. G. *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, 1973.

- [15] Shao, Jun. *Mathematical Statistics*, Springer-Verlag, New York, 2003.
- [16] Hochwald, Bertrand; Nehrai, Arye. On identifiability and information-regularity in parameterized normal distributions. *Circuits, Systems, and Signal Processing* **1997**, 16 (1), 83–89.
- [17] Stoica, Petre; Marzetta, Thomas L. Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing* **2001**, 49 (1), 87–90.
- [18] Boyd, Stephen; Vandenberghe, Lieven. *Convex Optimization*, Cambridge University Press, 2004.
- [19] Kirkwood, James R. *An Introduction to Analysis*, PWS Publishing Company, 1995.
- [20] Haftka, Raphael T.; Gürdal, Zafer. *Elements of Structural Optimization* Kluwer Academic Publishers, 1992.
- [21] Moon, Todd K.; Stirling, Wynn C. *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, 2000.
- [22] Bertsekas, Dimitri P. *Nonlinear Programming*, Athena Scientific, 1995.
- [23] Goldstein, A. A. Convex programming in Hilbert space. *Bulletin of the American Mathematical Society* **1964** 70 (5), 709–710.
- [24] Franklin, Joel. N. *Matrix Theory*, Dover, 1993.
- [25] Moore, Terrence J.; Sadler, Brian M.; Kozick, Richard J. Regularity and strict identifiability in MIMO systems. *IEEE Transactions on Signal Processing* **2002**, 50 (8), 1831–1842.
- [26] van der Veen, Alle-Jan. Asymptotic properties of the algebraic constant modulus algorithm. *IEEE Transactions on Signal Processing* **2001**, 49 (8), 1796–1807.
- [27] Bertsekas, Dimitri P. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control* **1976**, AC-21 (2), 174–184.
- [28] Luenberger, David G. *Optimization by Vector Space Methods*, John Wiley & Sons, Inc., 1969.
- [29] Kale, B. K. On the solution of likelihood equations by iteration processes. The multiparametric case. *Biometrika* **1962**, 49 (3–4), 479–486.
- [30] Kale, B. K. On the solution of likelihood equations by iteration processes *Biometrika* **1961**, 48, 452–6.
- [31] Goldstein, A. A. On gradient projection *Allerton Conf.* 1974, 38–40.

- [32] Gafni, Eli M.; Bertsekas, Dimitri P. Convergence of a gradient projection method. *Laboratory for Information and Decision Systems Technical Report*, LIDS-P-1201, May 1982.
- [33] Apostol, Tom M. *Mathematical Analysis*, Addison Wesley, 1974.
- [34] Blatt, Doron; Hero, Alfred. Distributed maximum likelihood estimation for sensor networks. *IEEE ICASSP 2004* **2004**, 3, 929–932.

A. Equivalence of Optimality Conditions

This appendix shows the equivalence between Bertsekas' criteria of optimality locally in a convex set (22,27) and the method of Lagrange multipliers. Bertsekas defines $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ to be optimal in the convex set $\boldsymbol{\Theta}$ provided

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq 0 \quad (\text{A.1})$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The Lagrange method is optimal given the KKT conditions in equations 20 through 24.

Assume equation A.1. Then either

- (a) $\boldsymbol{\theta}^* \in \text{int}\boldsymbol{\Theta}$ (the interior of $\boldsymbol{\Theta}$), or
- (b) $\boldsymbol{\theta}^* \in \partial\boldsymbol{\Theta}$ (the boundary).

First, suppose (a) $\boldsymbol{\theta}^* \in \text{int}\boldsymbol{\Theta}$. Then $\boldsymbol{\theta}^*$ is in an open set $\boldsymbol{O} \subset \boldsymbol{\Theta}$. Let $\boldsymbol{\theta}'$ be a point sufficiently close to $\boldsymbol{\theta}^*$ so that both $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}'' = \boldsymbol{\theta}^* - (\boldsymbol{\theta}' - \boldsymbol{\theta}^*)$ are in \boldsymbol{O} . Note that $\boldsymbol{\theta}''$ is the point vector exactly opposite $\boldsymbol{\theta}'$ from $\boldsymbol{\theta}^*$. Since $\boldsymbol{\theta}^*$ is optimal, then

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq 0.$$

for $\boldsymbol{\theta} = \boldsymbol{\theta}'$ and $\boldsymbol{\theta} = \boldsymbol{\theta}''$. But also

$$\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta}'' - \boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*) \cdot -(\boldsymbol{\theta}' - \boldsymbol{\theta}^*) \geq 0.$$

So, $\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = 0$ for every $\boldsymbol{\theta} \in \boldsymbol{O}$ since the choice of $\boldsymbol{\theta}'$ was arbitrary. And since \boldsymbol{O} is open the dimension of $\{\boldsymbol{\theta} - \boldsymbol{\theta}^* : \boldsymbol{\theta} \in \boldsymbol{O}\}$ is the dimension of $\boldsymbol{\Theta}$. Thus, we must have that $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{0}$. Simply choose $\boldsymbol{\mu}^* = \mathbf{0}$ and $\boldsymbol{\nu}^* = \mathbf{0}$, then the equation in (24) is satisfied by

$$\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*) + \boldsymbol{\mu}^{*T} \mathbf{F}(\boldsymbol{\theta}^*) + \boldsymbol{\nu}^{*T} \mathbf{G}(\boldsymbol{\theta}^*) = \mathbf{0}.$$

Otherwise, suppose (b) that $\boldsymbol{\theta}^* \in \partial\boldsymbol{Q}$ and suppose $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*) \neq \mathbf{0}$. Then since we have $f(\boldsymbol{\theta}) = \mathbf{0}$, $g(\boldsymbol{\theta}) \leq \mathbf{0}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $f(\boldsymbol{\theta}) = \mathbf{0}$, $g(\boldsymbol{\theta}) \geq \mathbf{0}$ for $\boldsymbol{\theta} \notin \boldsymbol{\Theta}$, and $\mathbf{F}(\boldsymbol{\theta}^*)$, $-\mathbf{G}(\boldsymbol{\theta}^*)$ must span the convex set locally at $\boldsymbol{\theta}^*$, then, in particular, there exists a $\boldsymbol{\mu}$, $\boldsymbol{\nu}$ with $\boldsymbol{\nu} \geq \mathbf{0}$ such that

$$\boldsymbol{\mu}^T \mathbf{F}(\boldsymbol{\theta}^*) - \boldsymbol{\nu}^T \mathbf{G}(\boldsymbol{\theta}^*) = -\epsilon \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}^*)$$

for some small $\epsilon > 0$. This is because the gradient of the negative log-likelihood must be in the direction of the convex set, otherwise $\boldsymbol{\theta}^*$ is not optimal. Therefore, $\boldsymbol{\mu}^* = -\epsilon^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\nu}^* = \epsilon^{-1}\boldsymbol{\nu}$ gives the desired KKT condition equation 24. (Note this requires the inequality constraint - for the strict equality constraint the set is not necessarily convex!)

Conversely, suppose we have the KKT condition equation 24. If the gradient of the log-likelihood is zero at a stationary point, then that stationary point is in $\text{int}\boldsymbol{\Theta}$ and the Bertsekas condition is trivial. So suppose it is nonzero and thus our stationary point $\boldsymbol{\theta}^*$ is on the boundary $\partial\boldsymbol{\Theta}$. Note $\mathbf{g}(\boldsymbol{\theta}^*) = \mathbf{0}$ on $\partial\boldsymbol{\Theta}$. Recall the Lagrangian in equation 19. Then note that for any $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$ we have that

$$\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) &= L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) + \nabla_{\boldsymbol{\theta}}^T L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(1) \\
&= L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) - \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\quad + \boldsymbol{\mu}^{*T} \mathbf{F}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \boldsymbol{\nu}^{*T} \mathbf{G}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + o(1) \\
&= L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) - \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\quad + \boldsymbol{\mu}^{*T} (\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\theta}^*)) + \boldsymbol{\nu}^{*T} (\mathbf{g}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta}^*)) + o(1) \\
&= L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) - \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \boldsymbol{\nu}^{*T} \mathbf{g}(\boldsymbol{\theta}) + o(1).
\end{aligned}$$

Since, $\boldsymbol{\theta}^*$ is optimal then $L(\boldsymbol{\theta}, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*) \geq L(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$, provided we scale $\boldsymbol{\nu}^*$ to be sufficiently small. Thus, $-\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq -\boldsymbol{\nu}^{*T} \mathbf{g}(\boldsymbol{\theta}) + o(1)$. Then $\boldsymbol{\theta}$ can be made arbitrarily close to $\boldsymbol{\theta}^*$ to force $o(1) \rightarrow 0$ and $\boldsymbol{\nu}^*$ can be scaled arbitrarily small and we have

$$-\nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \geq 0.$$

B. Taylor Expansion Derivation

Given the general iterative form equation 33, we derive the CSA via a Taylor expansion. Note that for any iterate $\boldsymbol{\theta}_k$ satisfying the constraints, the space spanned by the gradient of the active constraints at $\boldsymbol{\theta}_k$, namely $\mathbf{F}(\boldsymbol{\theta}_k)$, are the directions for steepest descent and ascent of \mathbf{f} . Thus, for the step \mathbf{d}_k to be feasible (or nearly so), it is necessary that \mathbf{d}_k be orthogonal (or nearly so) to this gradient matrix, i.e., $\mathbf{d}_k \in \text{span } \mathbf{U}(\boldsymbol{\theta}_k)$. Let \mathbf{d}'_k be such that $\mathbf{d}_k = \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k$ where $\mathbf{U}(\boldsymbol{\theta}_k)$ is defined in equation 5. So, the iterative form is now

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \cdot \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k. \quad (\text{B.1})$$

Now, consider the Taylor expansion of the log-likelihood about $\boldsymbol{\theta}_k$ along the vector $\mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k$. This is given by

$$\begin{aligned} \log p(\mathbf{x}; \boldsymbol{\theta}_k + \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k) &= \log p(\mathbf{x}; \boldsymbol{\theta}_k) + \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k \\ &\quad + \frac{1}{2} \mathbf{d}'_k{}^T \mathbf{U}^T(\boldsymbol{\theta}_k) \nabla_{\boldsymbol{\theta}}^2 \log p(\mathbf{x}; \boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k + o(\|\mathbf{d}'_k\|) \\ &\approx \log p(\mathbf{x}; \boldsymbol{\theta}_k) + \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k \\ &\quad - \frac{1}{2} \mathbf{d}'_k{}^T \mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k. \end{aligned}$$

Note above we replace the negative Hessian with the Fisher Information as well as drop the higher order terms. This gives, approximately, a quadratic log-likelihood model restricted to the subspace spanned by the columns of $\mathbf{U}(\boldsymbol{\theta}_k)$, i.e., the space which locally preserves the active constraints $\mathbf{f}(\boldsymbol{\theta}_k) = \mathbf{0}$, which is why it is referred to as the *null* space. The maximum of this quadratic model

$$\Phi(\mathbf{d}'_k) = \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k - \frac{1}{2} \mathbf{d}'_k{}^T \mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k \quad (\text{B.2})$$

over the choice of \mathbf{d}'_k must be the stationary point of the approximation. Hence, differentiating equation B.2 and setting the result to $\mathbf{0}$, the optimal \mathbf{d}'_k^* must satisfy

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}}^T \log p(\mathbf{x}; \boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k) - \mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)\mathbf{d}'_k^*.$$

This requires more than just the non-singularity of the $\mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)$ matrix, since the Hessian of the log-likelihood approximation must be negative definite as well for \mathbf{d}'_k^* to maximize this function. Indeed, in this case, the Hessian of $\Phi(\mathbf{d}'_k)$ is $-\mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)$, so we must have that $\mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k)$ is positive definite as well. Solving for \mathbf{d}'_k^* and substituting into \mathbf{d}_k in equation B.1 we obtain the unrestored version of the CSA,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha_k \mathbf{U}(\boldsymbol{\theta}_k) (\mathbf{U}^T(\boldsymbol{\theta}_k) \mathbf{I}(\boldsymbol{\theta}_k) \mathbf{U}(\boldsymbol{\theta}_k))^{-1} \mathbf{U}^T(\boldsymbol{\theta}_k) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}_k).$$

C. Constrained Least Squares Estimator

We will show that the CLSE in equation 65 is both unbiased and efficient relative to the Marzetta form of the CCRB under the assumption that $\mathbf{x} \sim \mathcal{N}(\mathbf{H}\boldsymbol{\theta}_o, \mathbf{C})$. We only show this for the real parameter case, although the complex parameter case is similar. We do so without the nullspace derivation of the CLSE. The equality shown in Theorem 3 is an alternative proof of this fact. Recall that the MLE $\bar{\boldsymbol{\theta}}(\mathbf{x})$ is unbiased, thus

$$\begin{aligned}
E_{\boldsymbol{\theta}_o} \hat{\boldsymbol{\theta}}_{CLSE}(\mathbf{x}) &= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} E_{\boldsymbol{\theta}_o} \bar{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \\
&= \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} (\mathbf{F} \boldsymbol{\theta}_o + \mathbf{v}) \\
&= \boldsymbol{\theta}_o.
\end{aligned}$$

So the CLSE is also unbiased. The covariance matrix of the CLSE is given by

$$\begin{aligned}
Cov_{\boldsymbol{\theta}_o}(\hat{\boldsymbol{\theta}}_{CLSE}(\mathbf{x})) &= E_{\boldsymbol{\theta}_o} \hat{\boldsymbol{\theta}}_{CLSE}(\mathbf{x}) \hat{\boldsymbol{\theta}}_{CLSE}^T(\mathbf{x}) - \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} E_{\boldsymbol{\theta}_o} (\bar{\boldsymbol{\theta}}(\mathbf{x}) \bar{\boldsymbol{\theta}}^T(\mathbf{x})) \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&\quad - \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} E_{\boldsymbol{\theta}_o} (\bar{\boldsymbol{\theta}}(\mathbf{x})) \mathbf{v}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \\
&\quad - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} E_{\boldsymbol{\theta}_o} (\bar{\boldsymbol{\theta}}^T(\mathbf{x})) \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&\quad + \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \mathbf{v}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} - \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} E_{\boldsymbol{\theta}_o} (\bar{\boldsymbol{\theta}}(\mathbf{x}) \bar{\boldsymbol{\theta}}^T(\mathbf{x})) \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&\quad - \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o \mathbf{v}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \\
&\quad - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \boldsymbol{\theta}_o^T \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&\quad + \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \mathbf{v}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} - \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} E_{\boldsymbol{\theta}_o} (\bar{\boldsymbol{\theta}}(\mathbf{x}) \bar{\boldsymbol{\theta}}^T(\mathbf{x})) \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&\quad + \left[\left(\left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \right) \right. \\
&\quad \left. \left(\left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \right)^T \right] \\
&\quad - \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) - \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T.
\end{aligned}$$

But since $\mathbf{F}\boldsymbol{\theta}_o + \mathbf{v} = \mathbf{0}$, then

$$\begin{aligned}
& \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{v} \\
&= \boldsymbol{\theta}_o - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} (\mathbf{F} \boldsymbol{\theta}_o + \mathbf{v}) \\
&= \boldsymbol{\theta}_o.
\end{aligned}$$

So this, with the knowledge that the MLE $\bar{\boldsymbol{\theta}}(\mathbf{x})$ is efficient relative to the CRB, shows that the covariance matrix of the CLSE is

$$\begin{aligned}
& \text{Cov}_{\boldsymbol{\theta}_o}(\hat{\boldsymbol{\theta}}_{CLSE}(\mathbf{x})) \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} [E_{\boldsymbol{\theta}_o}(\bar{\boldsymbol{\theta}}(\mathbf{x}) \bar{\boldsymbol{\theta}}^T(\mathbf{x})) - \boldsymbol{\theta}_o \boldsymbol{\theta}_o^T] \\
&\quad \times \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&= \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \mathbf{I} \left(\mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1} \right) \\
&= \mathbf{I}^{-1} - \mathbf{I}^{-1} \mathbf{F}^T (\mathbf{F} \mathbf{I}^{-1} \mathbf{F}^T)^{-1} \mathbf{F} \mathbf{I}^{-1}
\end{aligned}$$

which is the Marzetta form of the CCRB (2). Therefore the CLSE is both unbiased and efficient.

Distribution

ADMNSTR
DEFNS TECHL INFO CTR
ATTN DTIC-OCP (ELECTRONIC COPY)
8725 JOHN J KINGMAN RD STE 0944
FT BELVOIR VA 22060-6218

DARPA
ATTN IXO S WELBY
3701 N FAIRFAX DR
ARLINGTON VA 22203-1714

OFC OF THE SECY OF DEFNS
ATTN ODDRE (R&AT)
THE PENTAGON
WASHINGTON DC 20301-3080

US ARMY TRADOC
BATTLE LAB INTEGRATION & TECHL
DIRCTRT
ATTN ATCD-B
10 WHISTLER LANE
FT MONROE VA 23651-5850

SMC/GPA
2420 VELA WAY STE 1866
EL SEGUNDO CA 90245-4659

US ARMY ARDEC
ATTN AMSTA-AR-TD
BLDG 1
PICATINNY ARSENAL NJ 07806-5000

COMMANDING GENERAL
US ARMY AVN & MIS CMND
ATTN AMSAM-RD W C MCCORKLE
REDSTONE ARSENAL AL 35898-5000

US ARMY INFO SYS ENGRG CMND
ATTN AMSEL-IE-TD F JENIA
FT HUACHUCA AZ 85613-5300

US ARMY SIMULATION TRAIN &
INSTRMNTN CMND
ATTN AMSTI-CG M MACEDONIA
12350 RESEARCH PARKWAY
ORLANDO FL 32826-3726

US GOVERNMENT PRINT OFF
DEPOSITORY RECEIVING SECTION
ATTN MAIL STOP IDAD J TATE
732 NORTH CAPITOL ST., NW
WASHINGTON DC 20402

US ARMY RSRCH LAB
ATTN AMSRD-ARL-CI-OK-TP TECHL
LIB T LANDFRIED (2 COPIES)
ABERDEEN PROVING GROUND MD
21005-5066

DIRECTOR
US ARMY RSRCH LAB
ATTN AMSRD-ARL-RO-EV W D BACH
PO BOX 12211
RESEARCH TRIANGLE PARK NC 27709

US ARMY RSRCH LAB
ATTN AMSRD-ARL-CI J GOWENS
ATTN AMSRD-ARL-CI-C W INGRAM
ATTN AMSRD-ARL-CI-CN A SWAMI
ATTN AMSRD-ARL-CI-CN B SADLER
ATTN AMSRD-ARL-CI-CN G RACINE
(2 COPIES)
ATTN AMSRD-ARL-CI-CN S MISRA
ATTN AMSRD-ARL-CI-OK-T TECHL
PUB (2 COPIES)
ATTN AMSRD-ARL-CI-OK-TL TECHL
LIB (2 COPIES)
ATTN AMSRD-ARL-D J M MILLER
ATTN AMSRL-CI-CN T MOORE
(5 COPIES)
ATTN IMNE-ALC-IMS MAIL &
RECORDS MGMT
ADELPHI MD 20783-1197

INTENTIONELLY LEFT BLANK.